Context Aware Standard Process for Data Mining

http://www.casp-dm.org

Fernando Martínez-Plumed[1], Lidia Contreras-Ochando[1], Cèsar Ferri[1], Peter Flach[2], José Hernández-Orallo[1], Meelis Kull[3], Nicolas Lachiche[4] and María José Ramírez-Quintana[1]

[1]Departament de Sistemes Informàtics i Computació, Universitat Politècnica de València, Spain.
{fmartinez,cferri,jorallo,mramirez}@dsic.upv.es
[2]University of Bristol, U.K. peter.flach@bristol.ac.uk
[3]University of Tartu, Estonia. meelis.kull@ut.ee
[4]ICube, Université de Strasbourg, France. nicolas.lachiche@unistra.fr

Abstract

We propose an extension of the Cross Industry Standard Process for Data Mining (CRISP-DM) which addresses specific challenges of machine learning and data mining for context and model reuse handling. This new general context-aware process model is mapped with CRISP-DM reference model proposing some new or enhanced outputs.

Keywords: data mining, reframing, context awareness, process model, methodology.

1 Introduction

Anticipating potential changes in context is a critically important part of data mining projects. Unforeseen context changes can lead to substantial additional costs and in the extreme case require running a new project from scratch. For example, an automatic text summarisation system developed in the context of the English language can be extremely hard to be modified for other languages, unless such context change is anticipated. For another example, a fraud detection service provider develops its detectors in the context of known types of frauds, but the context keeps changing, with new types invented continuously. A careful analysis can help to build more versatile detectors which are effective for some new types of frauds and are easy to update for other new types. As a third example, a customer segmentation system helping to tailor products for different customer groups might be hard to modify to incorporate richer customer information, unless such context changes are anticipated.

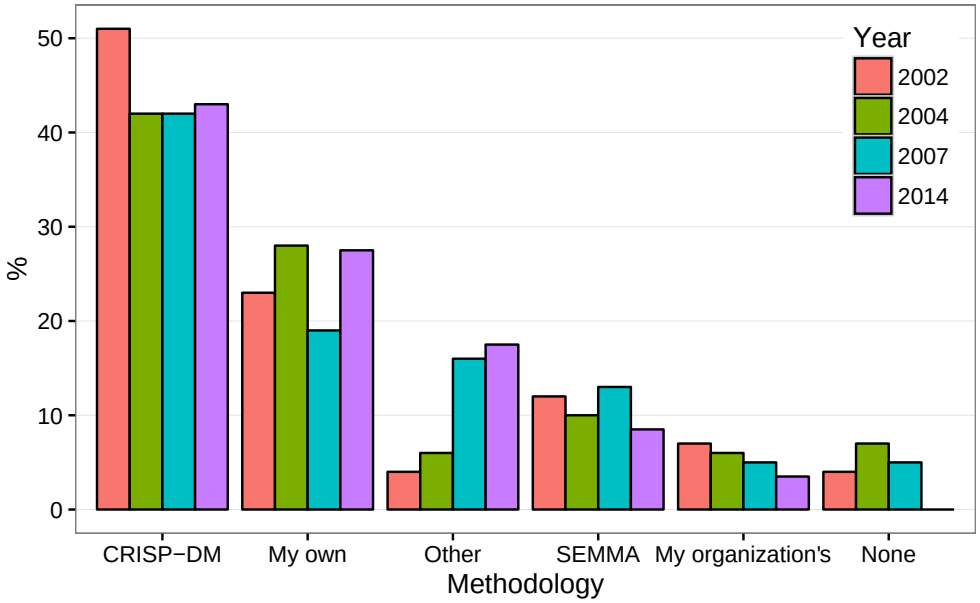

Figure 1: Use of data mining methodologies according to `www.kdnuggets.com`

Context anticipation is more than just a single separate task and it requires dedicated activities in all phases of the data mining process, from the initial domain understanding up to the final deployment. These activities are not included in any of the existing Data Mining (DM) standard process methodologies, such as the *Knowledge Discovery in Databases* (KDD) Process Fayyad et al. (1996a), the *Cross Industry Standard Process for Data Mining* (CRISP-DM) Chapman et al. (2000) and the *Sample, Explore, Modify, Model and Assess* (SEMMA)SAS (2005) process model. In this paper, we report on an extension of the CRISP-DM process model called CASP-DM (*Context-Aware Standard Process for Data Mining*), which has been evolving as a new standard with the goal of integrating context-awareness and context changes in the knowledge discovery process, while remaining backward compatible, so that users of CRISP-DM can adopt CASP-DM easily.

The reasons why we have use CRISP-DM as a base are multiple. CRISP-DM is the most complete data mining methodology in terms of meeting the needs of industrial projects and has become the most widely used process for DM projects, according to the KDnuggets polls held in 2002, 2004, 2007, and 2014. Although CRISP-DM does not seem to be maintained[1] or adapted to the new

[1] The original crisp-dm.org site is no longer active.

challenges in data mining, the proposed six phases and their subphases are still a good guide for the knowledge discovery process. In fact, the interest in CRISP-DM continues to be high compared to other models (see Figures 1 and 2). Therefore, the participation and cooperation of the data mining community is, of course, pivotal to the success of CASP-DM. This inclusion should imply the development of a platform where the data mining community can have access to the standard, which otherwise has the risk of being diluted, while working as an embryo for a committee and stable working group for an evolving standard accommodating future challenges and evolution of the field. Furthermore, CRISP-DM is supported by several project management software tools, such as RapidMiner[2] and IBM SPSS Modeler[3]. The extension of CRISP-DM into CASP-DM allows data mining projects to become context-aware while keep using these tools.

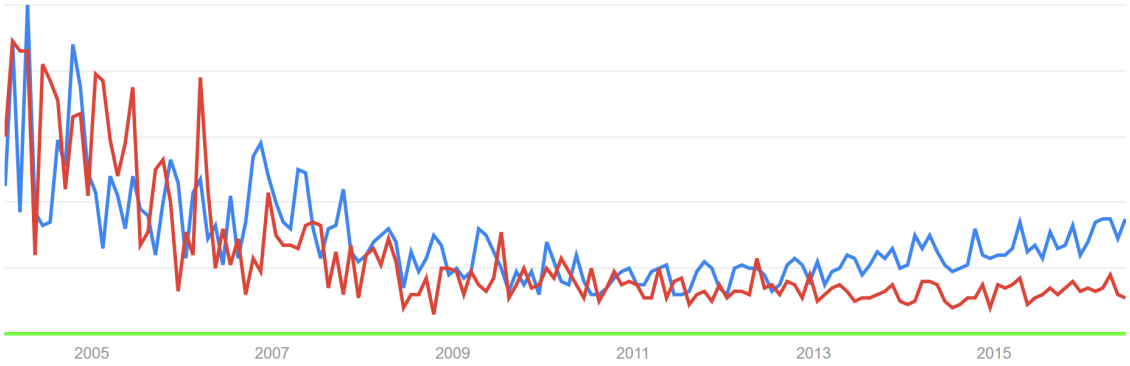

Figure 2: Relative interest over time in web searches according to Google Trends (`www.google.es/trends/`). Terms legend: CRISP-DM in blue, KDD in red, SEMMA in green (the latter having a relative interest close to zero).

The rest of the document is organised as follows. Section 2 briefly reviews CRISP-DM and related methodologies, and the state of the art in terms of standardisation and maintenance of the methology. Section 3 discusses the role that context (or domain) is having in DM applications and the main types of context and context changes (including changes in costs, data distribution and others). Section 4 proposes CASP-DM, with new tasks and outputs as well as enhancements to the original reference model thus allowing the practitioners to be aware of (and anticipate) the main types of context. Finally, section 5 closes the paper.

2 Review of DM and CRISP-DM methodologies

In this section we review the main approaches (process models and methodologies[4]) useful to extract useful information from large volumes of data (see Mariscal et al. (2010) for a complete survey). We focus on two main approaches: *Knowledge Discovery in Databases* (KDD) (Fayyad et al., 1996a,b) since it was the original approach, and the CRISP-DM (Chapman et al., 2000), since it is the reference methodology. The rest of approaches detailed are based on them. Figure 3 shows a diagram of how the different DM and KD process models and methodologies have evolved. Furthermore, Table 5 compares the phases into which the DM and KD process is decomposed according the proposals discussed.

[2] `https://rapidminer.com`

[3] `http://www.ibm.com/software/analytics/spss/products/modeler`

[4] While a *process model* is defined as a set of tasks to be performed to develop a particular element (as well as their inputs and outputs), a *methodology* can be defined as a process model instance, in which not only tasks, inputs and

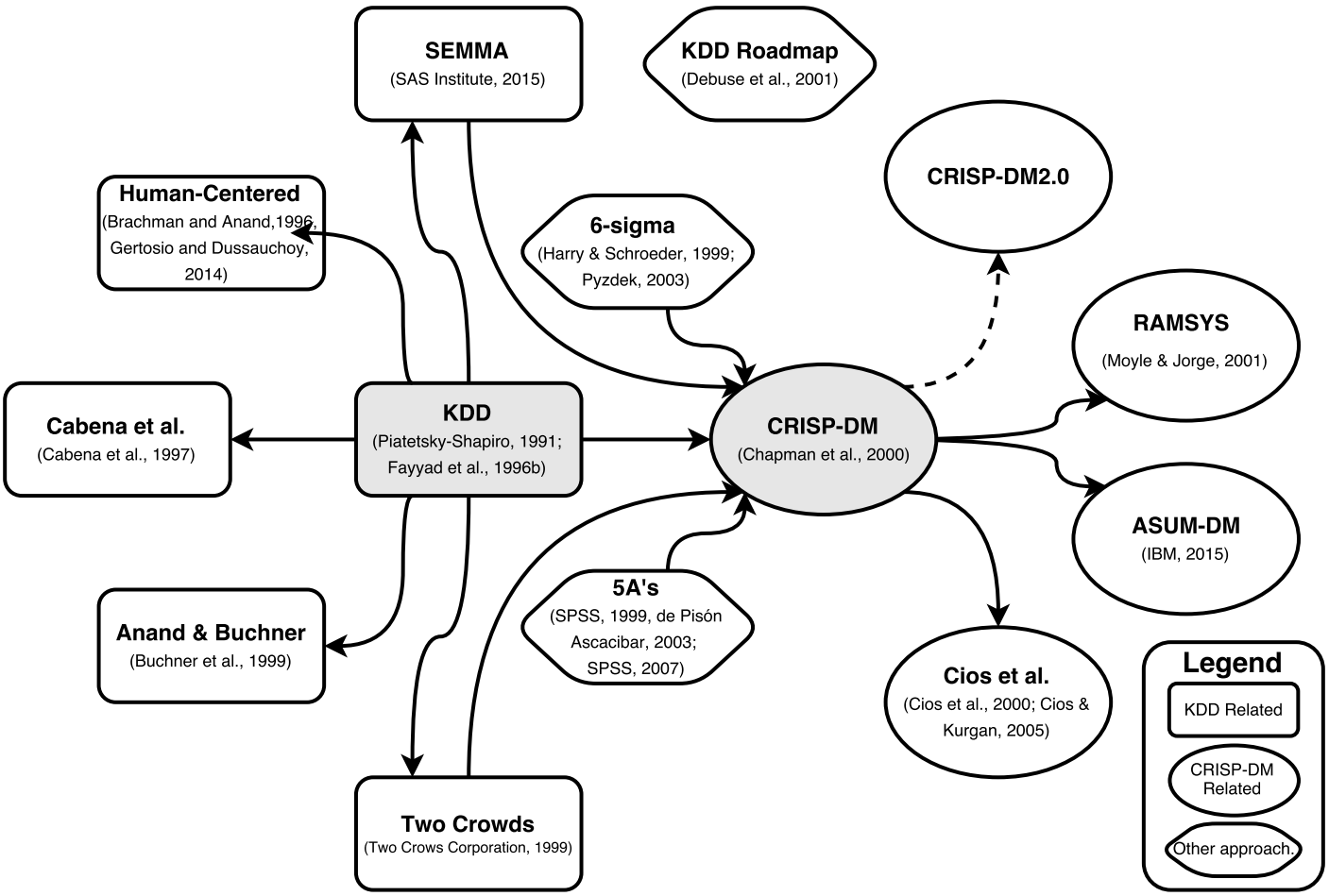

Figure 3: Evolution of DM Methodologies. Adapted from (Mariscal et al., 2010)

2.1 KDD related approaches

The term *Knowledge Discovery in Databases* (KDD) (Fayyad et al., 1996a,b) was the first process model to establish all the steps to be taken to develop a Data Mining project. According to Fayyad et al. Fayyad et al. (1996a) KDD is defined as "[...] the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." The non-trivial[5] goal is thus to (automatically) extract high-level knowledge that may not be obvious but potentially useful from raw or unprocessed data. This discovery of knowledge from a set of facts is accomplished by applying *Data Mining* (DM) methods. However, KDD has a much broader scope, of which DM is just one step in the whole process model. This process model involves several steps, including data processing, search for patterns, knowledge evaluation and interpretation, and refinement, where the whole process is interactive and iterative, which means that sometimes it may be necessary to repeat the previous steps. The overall process involves the repeated application of the following nine steps:

- **Developing an understanding of the application domain**, the relevant prior knowledge and the goals of the end-user.
- **Creating a target data set**: selecting a data set, or focusing on a subset of variables, or data

outputs must be specified but also the way in which the tasks must be carried out.

[5]Involving search or inference.

samples, on which discovery is to be performed.

- **Data cleaning and preprocessing**: including basic operations for removing noise or outliers, collecting necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time sequence information and known changes.
- **Data reduction and projection**: including finding useful features to represent the data depending on the goal of the task, using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.
- **Choosing the data mining task**: deciding whether the goal of the KDD process is classification, regression, clustering, etc.
- **Choosing the data mining algorithm(s)**: selecting method(s) to be used for searching for patterns in the data, deciding which models and parameters may be appropriate and matching a particular data mining method with the overall criteria of the KDD process.
- **Data mining**: searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth.
- **Knowledge interpretation**: interpreting the discovered patterns.
- **Consolidating discovered knowledge**: incorporating the discovered knowledge into the performance systems.

The different phases in the KDD process are outlined in Figure 2.1 where we see a large amount of unnecessary loops between steps and a lack of business guidance.

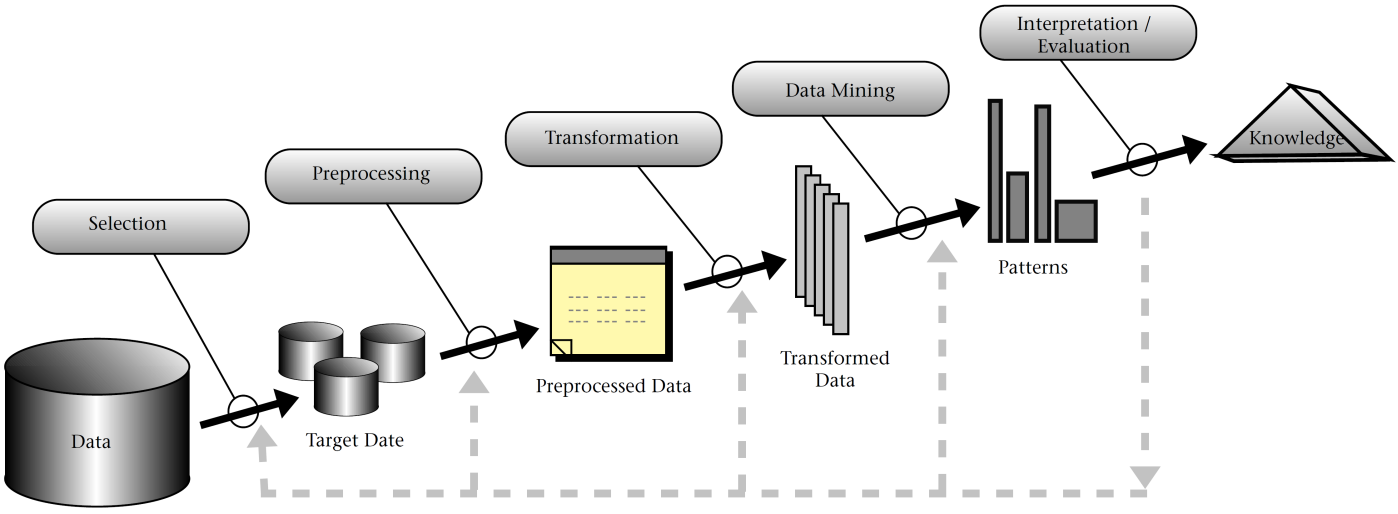

Figure 4: An Overview of the steps of the KDD Process (from Fayyad et al. (1996a))

Several other process models and methodologies have been developed using the KDD approaches as a basis. The *Human-Centered Approach to Data Mining* is presented in (Brachman and Anand, 1996; Gertosio and Dussauchoy, 2004). This proposal involves a holistic understanding of the

entire Knowledge Discovery Process and involves eight steps: human resource identification, problem specification, problem specification, data prospecting, methodology identification, data preprocessing, pattern discovery, and knowledge post-processing. It considers peopleâĂŹs involvement and interpretation in each process and put emphasis on that the target user is the data engineer.

SEMMA (SAS, 2005), which that stands for Sample, Explore, Modify, Model and Assess, is the methodology that the SAS institute[6] proposed for developing DM products. Although it is a methodology, it is based only on the technical part of the project and integrated into SAS tools such as *Enterprise Miner*. Unlike the former KDD process, SEMMA is not an open process and can only be used in these tools. The steps of SEMMA are mainly focussed on the modeling tasks of DM projects, leaving the business aspects. The steps are the following: sample, explore, modify, model and assess.

The two models by (Cabena et al., 1998) and (Anand and Büchner, 1998; Anand et al., 1998; Buchner et al., 1999) are based on KDD process with not big differences and with similar features. The former structures the process in a different number of steps (business objectives determination, selection, preprocessing and transformation, data mining, analysis of results and assimilation of knowledge) and was used more in the marketing and sales domain, this being one of the first process models which took into account the business objectives. For its part, the latter process model is adapted to web mining projects and focused on an online customer (incorporating the available operational and materialized data as well as marketing knowledge). The model consists of eight steps: human resource identification, problem specification, problem specification, data prospecting, methodology identification, data preprocessing, pattern discovery, and knowledge post-processing. Although it provides a detailed analysis for the initial steps, it does not include information on using the obtained knowledge.

The *Two Crows* Edelstein (1998) is a process model proposed by Two Crows Consulting[7] and takes advantage of some insights from (first versions of) CRISP-DM (before release). It proposes a non-linear list of steps (very close to the KDD phases), so it is necessary to go back and forth and . The basic steps of data mining for knowledge discovery are: define business problem, build data mining database, explore data, prepare data for modeling, build model, evaluate model, deploy model and results.

2.2 Independent approaches

There are some other independent approaches not related to the KDD original process. SPSS[8] originally developed a data mining analysis cycle called the *5 AâĂŹs Process* (Brunk et al., 1997) included their data mining tool set. It involves five steps (Assess, Access, Analyse, Act and Automate) where the "Automate" step is the most relevant one and helps non-experts user to automate the whole process of DM applying already defined methods to new data. The main disadvantage is that the 5 AâĂŹs do not contain steps to understand the business objectives and to test data quality. The process was abandoned in 1999 when SPSS joined CRISP-DM consortium to develop the CRISP-DM process model.

In mid-1996, Motorola developed the $6-\sigma$ approach (Harry, 1998) which emphasises measurement and statistical control techniques for quality and excellence in management. It is a well structured data-driven methodology for eliminating defects or quality control problems in manufacturing, service delivery, management, and other business activities, including data mining projects. That is done through the application of so-called "Six Sigma DMAIC" sequence of steps (Define,

[6] http://www.sas.com

[7] http://twocrows.com/

[8] http://www.spss.com.hk/

	Domain	#	Phases								
KDD	Academic	5	Selection	Pre processing	Transformation	Data Mining	Interpretation/ Evaluation				
KDD Fayyad	Academic	9	Developing and Understanding of the Application Domain	Creating a Target Data Set	Data Cleaning and Pre-processing	Data Reduction and Projection	Choosing the DM Task	Choosing the DM Algorithm	Data Mining	Interpreting Mined Patterns	Consolidating Discovered Knowledge
5 A's	Industry	5	Asses	Access	Analyse	Act	Automate				
6-sigma	Industry	5	Define	Measure	Analyse	Improve	Control				
Human Centered	Academic	6	Task Discovery	Data Discovery	Data Cleaning	Model Development	Data Analysis	Output Generation			
SEMMA	Industry	5	Sample	Explore	Modify	Model	Assess				
Two Crows	Industry	7	Define Business Problem	Build DM Data Base	Explore Data	Prepare Data for Modeling	Build Model	Evaluate Model	Deploy Model and Results		
Annand & Buchner	Academic	8	Domain Knowledge Elicitation	Human resource Identification	Problem Specification	Data Prospecting	Methodology Identification	Data Pre-processing	Pattern Discovery	Knowledge Post-processing	
Cabena	Industry	5	Select	Pre-process	Transform	Mining	Analyse and Assimilate				
Cios	Hybrid	6	Understanding the Problem Domain	Understanding the Data	Preparation of the Data	Build Model	Evaluation of the Discovered Knowledge	Using the Discovered Knowledge			
KDD Roadmap	Industry	8	Resourcing	Problem Specification	Data Cleansing	Pre-processing	Data Mining	Evaluation	Interpretation	Exploitation	
CRISP-DM	Industry	6	Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment			

Figure 5: Phases of Data Mining Methodologies

Measure, Analyze, Improve, and Control). This methodology has proven to be successful in companies such as IBM, Microsoft, General Electric, Texas Instrument or Ford.

KDD Roadmap (Debuse et al., 2001) is an iterative data mining methodology methodology used in Witness Miner toolkit[9] which uses a visual stream-based interface to represent routes through the KDD roadmap (consisting of eight steps: problem specification, resourcing, data cleansing, preprocessing, data mining, evaluation, interpretation and exploitation). The main contribution of KDD roadmap is the resourcing task which consist in the integration of databases from multiple sources to form the operational database.

2.3 CRISP-DM: de facto standard

We focus on *Cross Industry Standard Process for Data mining* (CRISP-DM) (Chapman et al., 2000) as a process model because it is the "de facto standard" for developing DM and KD projects. In addition, CRISP-DM is the most used methodology for developing DM projects[10]. In general terms, CRISP-DM is a general purpose process model which is a freely available, industry independent, technology neutral, and it is said to be de facto standard for DM.

CRISP-DM, as a process model, provides an overview of the life cycle of a data mining project. It contains the phases of a project, a set of tasks to be performed in each phase as well as the elements that are produced in each task (outputs) and the elements that are necessary to do a task (inputs). The life cycle of a data mining project consists of six phases (Figure 6) which sequence is not rigid: moving back and forth between different phases is always required and depends on the outcome of each phase which phase or which particular task of a phase, has to be performed next. The arrows indicate the most important and frequent dependencies between phases. The outer circle in Figure 6 symbolizes the cyclical nature of data mining itself. Data mining is not over once a solution is deployed. Therefore data mining processes will benefit from the experiences of previous ones.

In the following, we outline each phase briefly following the original reference model in (Chapman et al., 2000):

[9] http://www.witnessminer.com/

[10] CRISP-DM is still the top methodology for analytics, data mining, or data science projects according to *kDnuggets*: http://goo.gl/CYISan

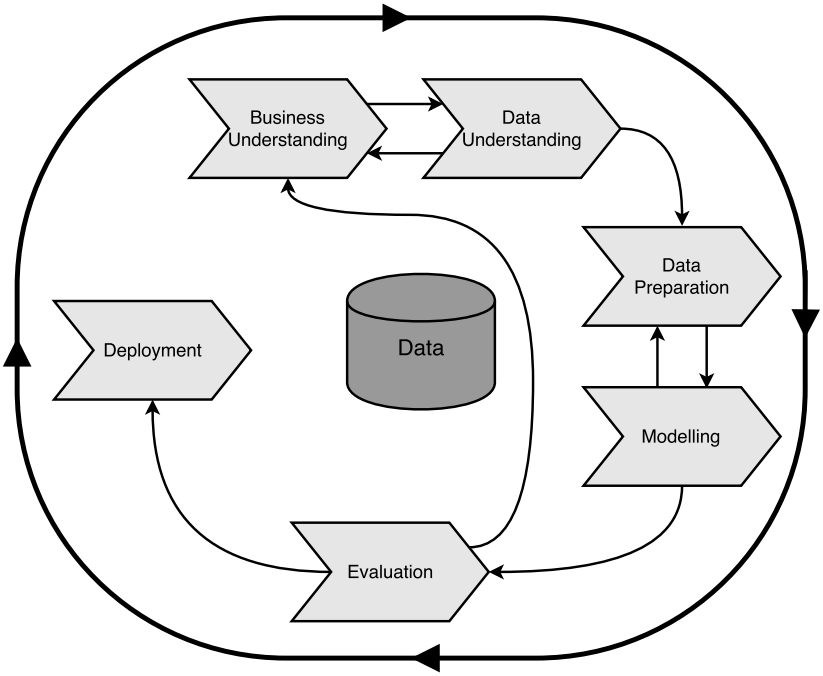

Figure 6: Process diagram showing the relationship between the different phases of CRISP-DM

1. **Business understanding**: This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

2. **Data understanding**: The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

3. **Data preparation**: The data preparation phase covers all activities to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record and attribute selection as well as transformation and cleaning of data for modeling tools.

4. **Modeling**: In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often necessary.

5. **Evaluation**: At this stage the model (or models) obtained are more thoroughly evaluated and the steps executed to construct the model are reviewed to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached

6. **Deployment**: Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organised and presented in a way that the customer can use it.

Its final goal is to make the process repeatable, manageable and measurable (to be able to get metrics). CRISP-DM is usually referred as an informal methodology (although it does not provide the rigid framework, task/inputs/outputs specification and execution, evaluation metrics, or correctness criteria) because it provides the most complete tool set for DM practitioners. The current version includes the reference process model and implementation user guide defining phases, tasks, activities and deliverable outputs of these tasks.

It is clear from Figure 3 that CRISP-DM is the standard model and has borrowed principles and ideas from the most important models (KDD, SEMMA, Two Crowds,...) and has been the source for many later proposals. However, many changes have occurred in the business application of data mining since the former version of CRISP-DM was published: new data types and data mining techniques and approaches, more demanding requirements for scalability, real-time deployment and large-scale databases, etc. The *CRISP-DM 2.0* Special Interest Group (SIG) was established with the aim of meeting the changing needs of DM with and improved version of the CRISP-DM process. Normally this version should have appeared in 2007, but was finally discontinued.

However, other process models based on the original CRISP-DM approach have appeared. Cios et al.'s six-step discovery process (Cios et al., 2000; Cios and Kurgan, 2005) was first proposed in 2000 adapting the CRISP-DM model to the needs of the academic research community. The main extensions include, among others, improved (research-oriented) description of the steps, explicit feedback mechanisms, reuse of knowledge discovered between different domains, etc. The model consists of six steps: understanding the problem domain, understanding the data, preparation of the data, data mining, evaluation of the discovered knowledge and using the discovered knowledge.

The *RAMSYS* (RApid collaborative data Mining SYStem) (Moyle and Jorge, 2001) is a methodology for developing DM and KD projects where several geographically diverse groups (nodes) work together on the same problem in a collaborative way. This methodology, although based on CRISP-DM (same phases and generic tasks), emphasises collaborative work, knowledge sharing and communication between groups. Apart from the original CRISP-DM tasks, the RAMSYS methodology proposes a new task called model submission (modeling step), where the best models from each of the nodes are evaluated and delivered.

Finally, in 2015, IBM Corporation released *ASUM-DM* (Analytics Solutions Unified Method for Data Mining/Predictive Analytics) a new methodology which refines and extends CRISP-DM. ASUM-DM retained the âĂIJAnalyticalâĂİ activities and tasks of CRISP-DM but the method was augmented adding infrastructure, operations, deployment and project management sections as well as templates and guidelines.

3 Context-awareness and reuse of knowledge

A major assumption in many machine learning and data mining algorithms is that the training and deployment data must be in the same contexts, namely, having the same feature space, distribution or misclassification cost. However, in many real-world applications, this assumption may not hold. Apart from having several different training contexts, there might also be many potential deployment contexts which differ from the training context(s) in one or more ways. An illustrative, nor exhaustive, list of context changes is shown is Table 1.

Many recent machine learning approaches have addressed the need to cope with context changes and reuse of learnt knowledge. Areas such as *data shift* Quiñonero-Candela et al. (2009); Moreno-Torres et al. (2012); Kull and Flach (2014), *domain adaptation* Jiang (2008), *transfer learning* Torrey and Shavlik (2009); Pan and Yang (2010), *transportability* Bareinboim and Pearl (2012), *meta-learning* Giraud-Carrier et al. (2004), *multi-task learning* Caruana (1997); Thrun (1996), *learning*

Context change	Examples of parametrised context
Distribution shift (covariate, prior probability, concept)	Input or output variable distribution
Costs and evaluation function	Cost proportion, cost matrix, loss function
Data quality (uncertain, missing, or noisy information)	Noise or uncertainty degree, missing attribute set
Representation change, constraints, background knowledge	Granularity level, complex aggregates, attribute set
Task change	Binarised regression cut-off, bins

Table 1: Taxonomy of context change types and examples of their parametrisation.

from noisy data Angluin and Laird (1988); Frénay and Verleysen (2013), *context-aware computing* Abowd et al. (1999), *mimetic models* Blanco-Vega et al. (2006), *theory revision* Richards and Mooney (1991), *lifelong learning* Thrun and Pratt (2012) and *incremental learning* Khreich et al. (2012). Generally, in these areas the context change is analysed when it happens, rather than being anticipated, thus learning a model in the new context and reusing knowledge from the original context.

A more proactive way to deal with context changes is by constructing a *versatile* model, which has the distinct advantage that it is not fitted to a particular context or context change, and thus enables model reuse. A new and generalised machine learning approach called *Reframing* Hernández-Orallo et al. (2016) addresses that. It formalises the expected context changes before any learning takes place, parametrises the space of contexts, analyses its distribution and creates versatile models that can systematically deal with that distribution of context changes. Therefore, the versatile model is reframed using the particular context information for each deployment situation, and not retrained or revised whenever the operating contexts change (see Figure 7). Rather than being an umbrella term for the above-mentioned related areas, reframing is a distinctive way of addressing context changes by anticipating them from the outset. *Cost-sensitive learning* Elkan (2001); Turney (2000); Chow (1970); Tortorella (2005); Pietraszek (2007); Vanderlooy et al. (2006) and *ROC analysis and cost plots* Metz (1978); Flach et al. (2003); Fawcett (2006); Flach (2010); Drummond and Holte (2006); Flach et al. (2011); Hernández-Orallo et al. (2011); Hernández-Orallo et al. (2012a); Hernández-Orallo et al. (2013) can be seen as areas where reframing has been commonly used in the past, and generally restricted to binary classification.

Generally speaking, the process of preparing a model to perform well over a range of different operation contexts involves a number of challenges:

- **Reuse of learnt knowledge**: Models are required to be more general and adaptable to changes in the data distribution, data representation, associated costs, noise, reliability, background knowledge, etc. This naturally leads to a perspective in which models are not continuously retrained and re-assessed every time a change happens, but rather kept, enriched and validated in a long-term model life-cycle. This lead us to the concept of versatile models, able to generalise over a range of contexts.

- **Variety of contexts and context changes**: The process of preparing and devising a versatile model to perform well over a range of operating contexts (beyond the specific context in which the model was trained) involves to deal with a number of different possible context changes that are commonly observed in machine learning applications: distribution shift Kull and Flach (2014); Moreno-Torres et al. (2012); Quiñonero-Candela et al. (2009), cost and evaluation function Elkan (2001); Turney (2000); Chow (1970); Pietraszek (2007); Tortorella (2005); Vanderlooy et al. (2006), data quality Frénay and Verleysen (2013), representation change Martínez-Usó and Hernández-Orallo (2015); Martínez-Usó et al. (2015), constrains, background knowledge, task change Scheirer et al. (2013); Hernández-Orallo et al. (2016),...

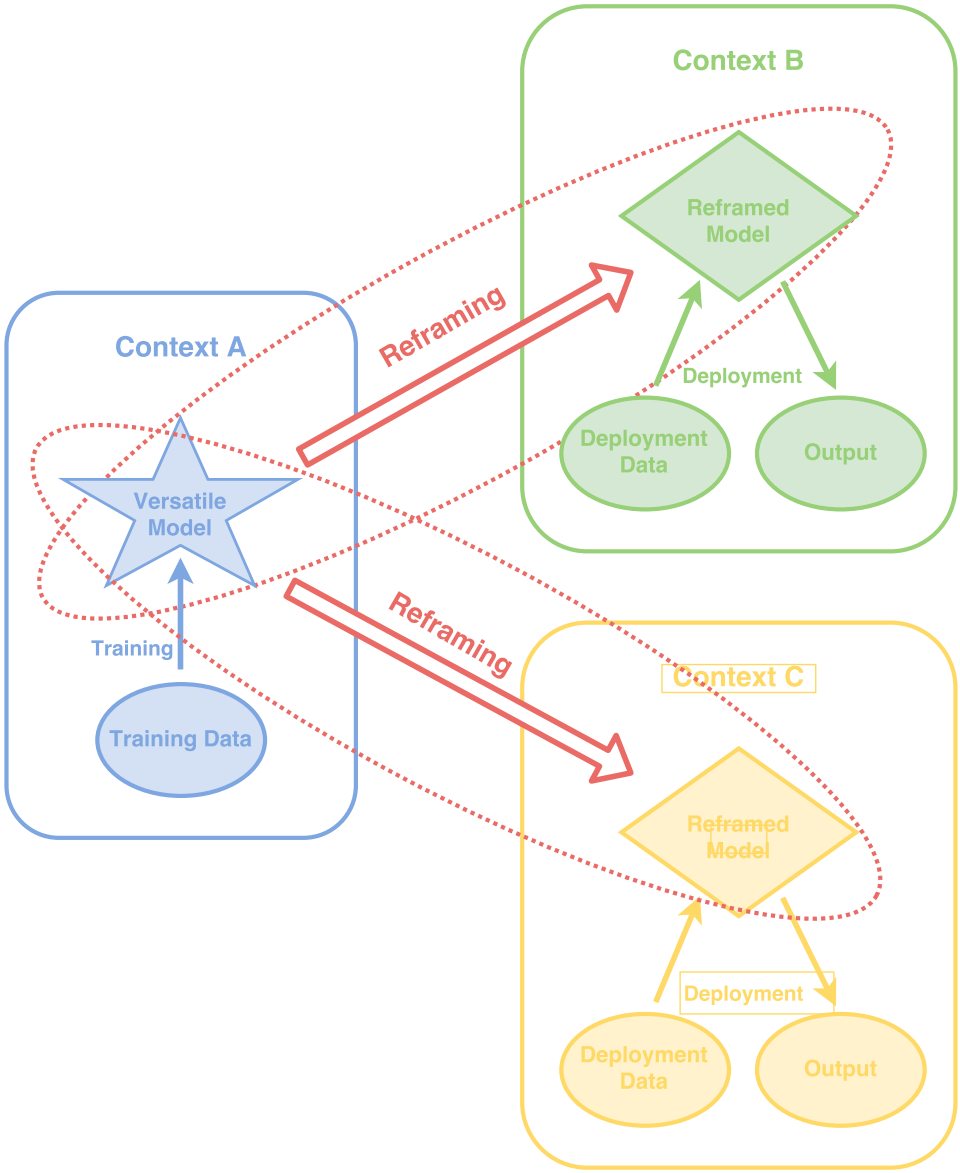

Figure 7: Operating contexts, models and reframing. The model on the left is intentionally more versatile than strictly necessary for context A, in order to ease its reframing to other contexts (e.g., B and C) without retraining it repeatedly.

- **Context-aware approaches for machine learning: Retraining vs. Revision vs. Reframing trilemma**: *Retraining* on the training data is very general, but there are many cases where it is not applicable. For instance, the training data may have been lost or may not exist (e.g., training models that have been created or modified by human experts) or may be prohibitively large (if deployment must work in restricted hardware), or the computational constraints do not allow retraining for each deployment context separately. Retraining on the deployment data can work well if there is an abundance of deployment data, but often the deployment data are limited, unsupervised or simply non-existent. A common alternative to retraining is *revision*, Raedt (1992); Richards and Mooney (1991) where parts of the model are patched or extended according to a new context (detection of novelty or inconsistency of the new data with respect to the existing model). It is especially natural as a result of an incremental learning Khreich et al. (2012) or lifelong learning Thrun and Pratt (2012). Finally, *reframing*, as

said above, is a context-aware approach that reuses the model trained in the training context by subjecting it to a reframing procedure that takes into account the particular deployment context .

- **Context-aware performance evaluation and visualisation**: When the context is constant, conventional context insensitive performance metrics can be used to evaluate how a model performs for that context. However, when we use the same model for several contexts we need context-aware performance metrics Ferri et al. (2009); Hernández-Orallo et al. (2011); Hernández-Orallo et al. (2013); Flach et al. (2011); Hand (2009); Hernández-Orallo et al. (2015, 2012b); Drummond and Holte (2006); Lo et al. (2011); Hernández-Orallo (2013); Xu et al. (2014); Kull and Hernández-Orallo (2015); Bi and Bennett (2003); Fawcett (2006); Flach (2010); Flach et al. (2003); Metz (1978).

These challenges require a change of methodology. If we have to be more anticipative with context, we need a process model where context is present from the very beginning, and the analysis, identification and use of context (changes) must be part of several stages. This is what CASP-DM undertakes.

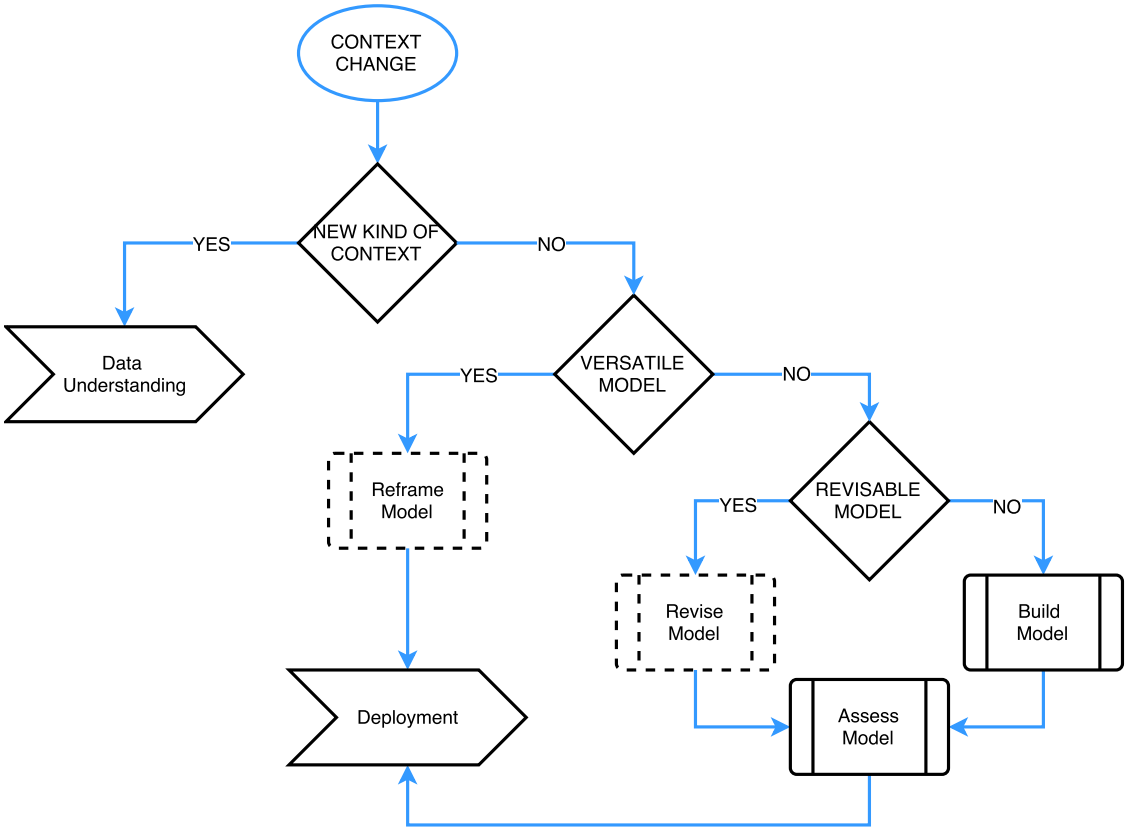

Figure 8: Tasks to be accomplished when there is a context change.

4 CASP-DM

CASP-DM, which stands for Context-Aware Standard Process for Data Mining, is the proposed extension of CRISP-DM for addressing specific challenges of machine learning and data mining for

context and model reuse handling. CASP-DM model inherits flexibility and versatility from the CRISP-DM life cycle and put more emphasis in that the sequence of phases is not rigid: context changes may affect different tasks so it should be possible to move to the appropriate phase. This is illustrated in Figures 8 (simplified) 9 (complete), where a flow chart shows which tasks in the CASP-DM process model should be completed whenever a context change needs to be addressed.

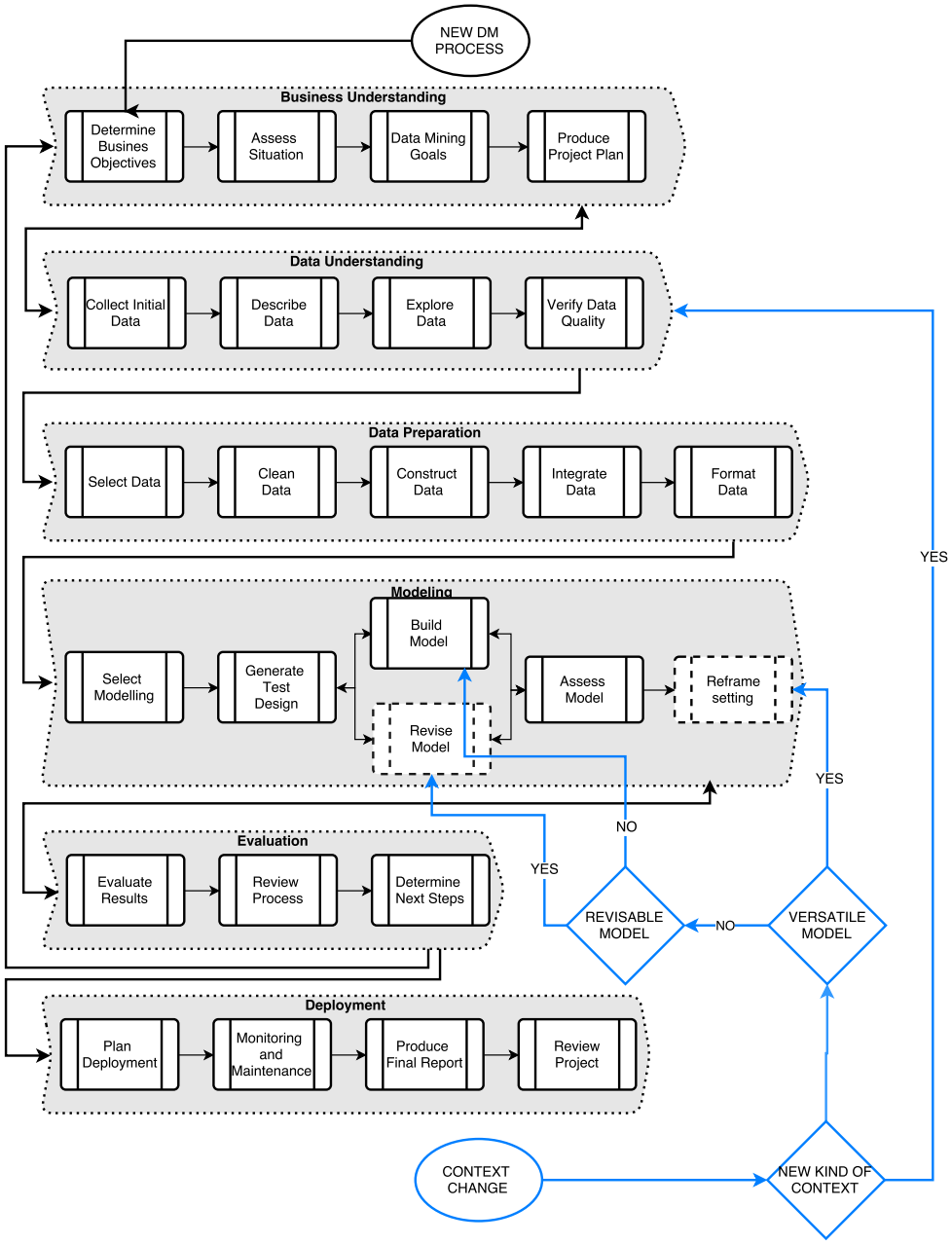

Figure 9: Complete view of the CASP-DM tasks to be completed whenever (1) a new context-aware DM project starts; or (2) a context change needs to be addressed.

In this section we overview the life cycle of a DM project by putting emphasis on those new and enhanced tasks and outputs that have to do with context and model reuse handling (Figure 10). Enhanced or new tasks/outputs are shown in **dark red**. Furthermore, a running example of model reuse with bike rental station data (MoreBikes) Kull et al. (2015b) will be used to illustrate how CASP-DM is applied in a real environment.

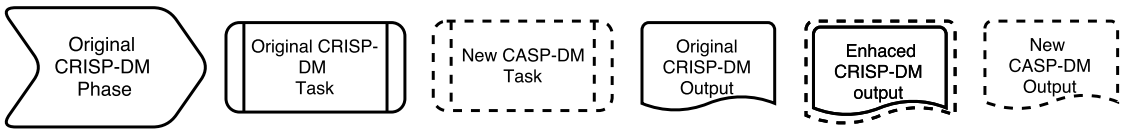

Figure 10: Legend of the different representation of original and new/enhanced tasks and outputs.

4.1 Business Understanding

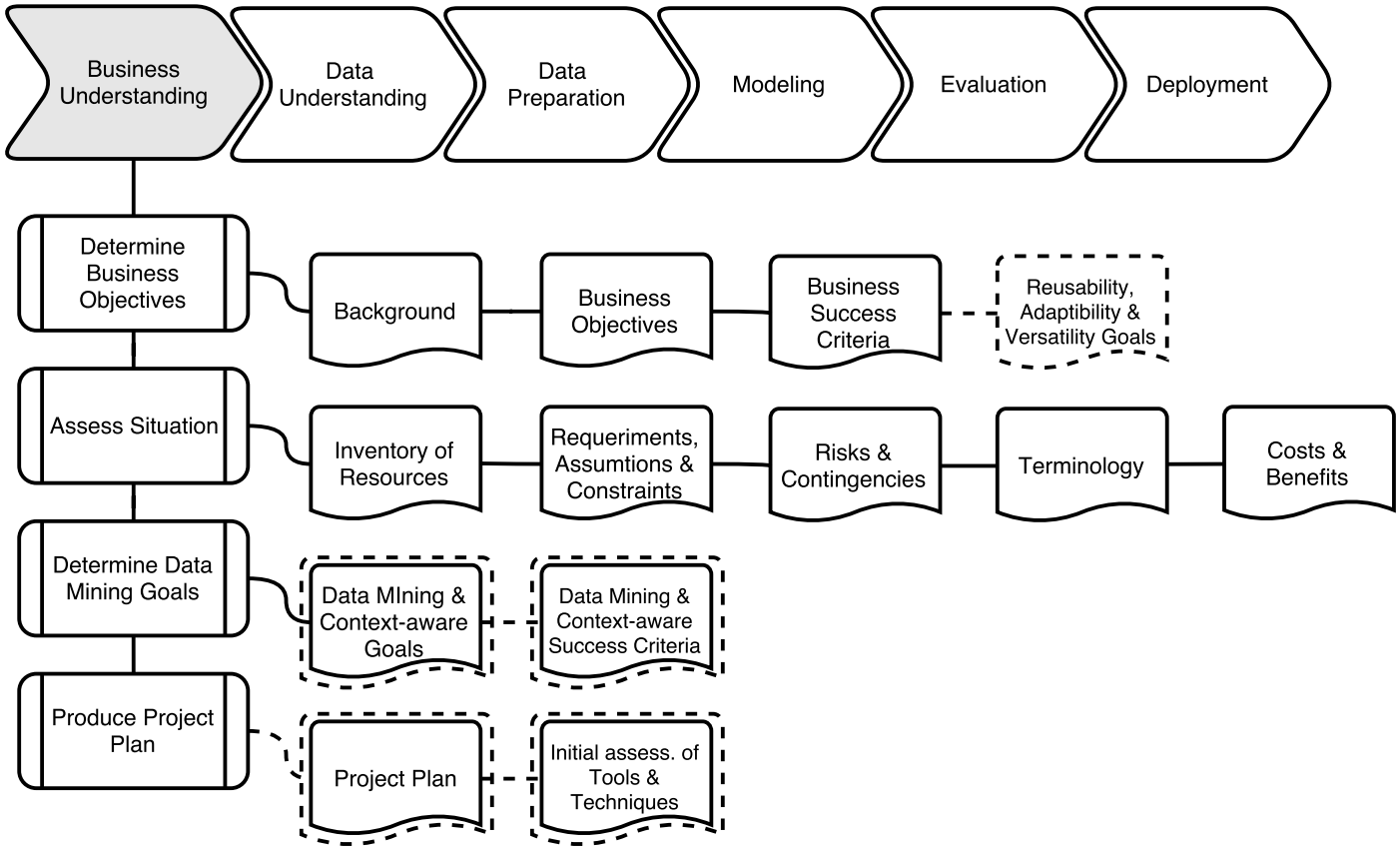

Figure 11: Phase 1. Business understanding: tasks and activities for context-awareness

The CASP-DM first phase “Business understanding” (as well as the second phase “Data understanding”) is the phase where the data mining project is being understood, defined and conceptualized. The rest phases are implementation-related phases, which aim to resolve the tasks being set in the first phases. As in the original CRISP-DM, the implementation phases are highly incremental and iterative where the lessons learned during the process and from the deployed solutions can benefit subsequent data mining processes.

The initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives. Adapting this phase to address context changes and model reuse handling involves: (1) adding new specialized tasks for identifying long term reusability business goals (whether the business goals involve reusability, adaptability, and versatility) w.r.t. context changes, (2) determining both data mining goals and success criteria when we address a context-aware data mining problem (which type of context-aware technique should be used depends on what aspects of the model are reusable in other contexts) and, finally, (3) perform an initial assessment of available context-aware techniques and update the project plan describing

the intended plan for achieving the data mining goals and thereby achieving the reusability, adaptability, and versatility business goals. The plan should specify the steps to be performed during the rest of the project, including the initial identification of contexts (changes), and the reframing techniques (I/O, structural) to deal with them.

4.1.1 Determine business objectives

- **Task**: The first task is to thoroughly understand, from a business perspective, what the client really wants to accomplish, and thus try to gain as much insight as possible into the business goals for data mining. For that it is necessary to gather background information about the current business situation, document specific business objectives and agree upon criteria used to determine data mining success from a business perspective.
- **Outputs**:
  - **Background**: Record the information that is known about the organizationâĂŹs business situation: determine organizational structure, identify the problem area and describe any solutions currently used to address the business problem
  - **Business objectives**: Describe the customerâĂŹs primary objective agreed upon by the project sponsors and other business units affected by the results
  - **Business success criteria**: Define the nature of business success for the data mining project from the business point of view. This might be as precisely as possible and able to be measured objectively.
  - **Reusability, Adaptability and Versatility Goals**: Identify, from a business long-term perspective, which are the prerequisites and future perspectives: whether the business goals involve reusability, adaptability, and versatility (i.e., should our solution procedure perform well over a range of different operating contexts?).

MoReBikeS example 1.

Finding Business Objectives
Adaptive reuse of learnt knowledge is of critical importance in the majority of knowledge-intensive application areas, particularly when the context in which the learnt model operates can be expected to vary from training to deployment. The MoReBikeS challenge (Model Reuse with Bike Rental Station Data) organised as the ECML-PKDD 2015 Discovery Challenge #1 Kull et al. (2015a), is focused on model reuse and context change.
The MoReBikeS challenge was carried out in the framework of historical bicycle rental data obtained from Valencia, Spain. Bicycles are continuously taken from and returned to rental stations across the city. Due to the patterns in demand some stations can become empty or full, such that more bikes cannot be rented or returned. To reduce the frequency of this happening, the rental company has to move bikes from full or nearly full stations to empty or nearly empty stations. Therefore the task is to predict the number of available bikes in every bike rental stations 3 hours in advance. There are at least two use cases for such predictions.

- First, a specific user plans to rent (or return) a bike in 3 hours time and wants to choose a bike station which is not empty (or full).
- Second, the company wants to avoid situations where a station is empty or full and therefore needs to move bikes between stations. For this purpose they need to know

which stations are more likely to be empty or full soon.

Context-awareness: Information from older stations should be used to improve performance on younger ones. In future, new stations will be planned every few months, but probably the growth is getting faster.

4.1.2 Assess situation

- **Task**: Once the goal is clearly defined, this task involves more detailed fact-finding about all of the resources, constraints, assumptions and other factors that should be considered in determining the data analysis goal and project plan.
- **Outputs**:
  - **Inventory of resources**: Accurate list of the resources available to the project, including: personnel, data sources, computing resources and software.
  - **Requirements, assumptions and constraints**: List all requirements of the project (schedule of completion, security and legal restrictions, quality, etc.), list the assumptions made by the project (economic factors, data quality assumptions, non-checkable assumptions about the business upon which the project rests, etc.) and list the constraints on the project (availability of resources, technological and logical constraints, etc.).
  - **Risks and contingencies**: List of the risks or events that might occur to delay the project or cause it to fail (scheduling, financial, data, results, etc.) and list of the corresponding contingency plans.
  - **Terminology**: Compile a glossary of technical terms (business and data mining terminology) and buzzwords that need clarification.
  - **Costs and benefits**: Construct a cost-benefit analysis for the project (comparing the estimated costs with the potential benefit to the business if it is successful).

MoReBikeS example 2.

Assessing the Situation
One of the first tasks the consultant faces is to assess the companyâĂŹs resources for data mining.

- **Data**. Since this is an established company, there is plenty of historical information from stations as well as information about the current status, time of the day/week/year, geographical data, weather conditions, etc.

4.1.3 Determine data mining goals

- **Task**: Translate business goals (in business terminology) into data mining goal reality (in technical terms).
- **Outputs**:

- **Data mining and context-aware goals**: Describe the type of data mining problem. Initial exploration of how the different contexts are going to be used. Describe technical goals. Describe the desired outputs of the project that enables the achievement of the business objectives.
- **Data mining and context-aware success criteria**: Define the criteria for a successful outcome to the project in technical terms: describe the methods for model and context assessment, benchmarks, subjective measurements, etc.

MoReBikeS example 3.

Data Mining Goals
Bike rental company needs to move bikes around to avoid empty and full stations. This can be done more efficiently if the numbers of bikes in the stations are predicted some hours in advance. The quality of such predictions relies heavily on the recorded usage over long periods of time. Therefore, the prediction quality on newly opened stations is necessarily lower. The goals for the study are:

- Use historical information about bike availability in the stations. In this challenge we explore a setting where there are 200 stations which have been running for more than 2 years and 75 stations which have just been open for a month.
- Reuse the models learned on 200 "old" stations in order to improve prediction performance on the 75 "new" stations. Combine information from similar stations to build improved models. Hence, this challenge evaluates prediction performance on the 75 stations.
- By predicting the number of bikes in the new stations (3 hours in advance), the bike rental company will be able to move bikes around to avoid empty and full stations.

4.1.4 Produce project plan

- **Task**: Describe the intended plan for achieving the data mining goals and thereby achieving the business goals. The plan should specify he project of the business goals, data mining goals (reusability, adaptability, and versatility), resources, risks, and schedule for all phases of data mining as well as include an initial selection of tools and techniques.
- **Outputs**:
  - **Project plan**: List the stages to be executed in the project, together with duration, resources required, inputs, outputs and dependencies. Where possible make explicit the large-scale iterations in the data mining process, for example repetitions of the modeling and evaluation phases.
  - **Initial assessment of tools and techniques**: At the end of the first phase, the project also performs an initial assessment of tools and techniques, including the initial identification of contexts (changes) and the context-aware techniques to deal with them.

MoReBikeS example 4.

MoReBikeS Example—Assessing Tools and Techniques
After setting the project plan for the study, an initial selection of tools and techniques should be made taking into account contexts and context changes:

- In this challenge, context is the combination of station and time. It should be advisable to use model combination, and retraining on sets of similar station.

4.2 Data Understanding

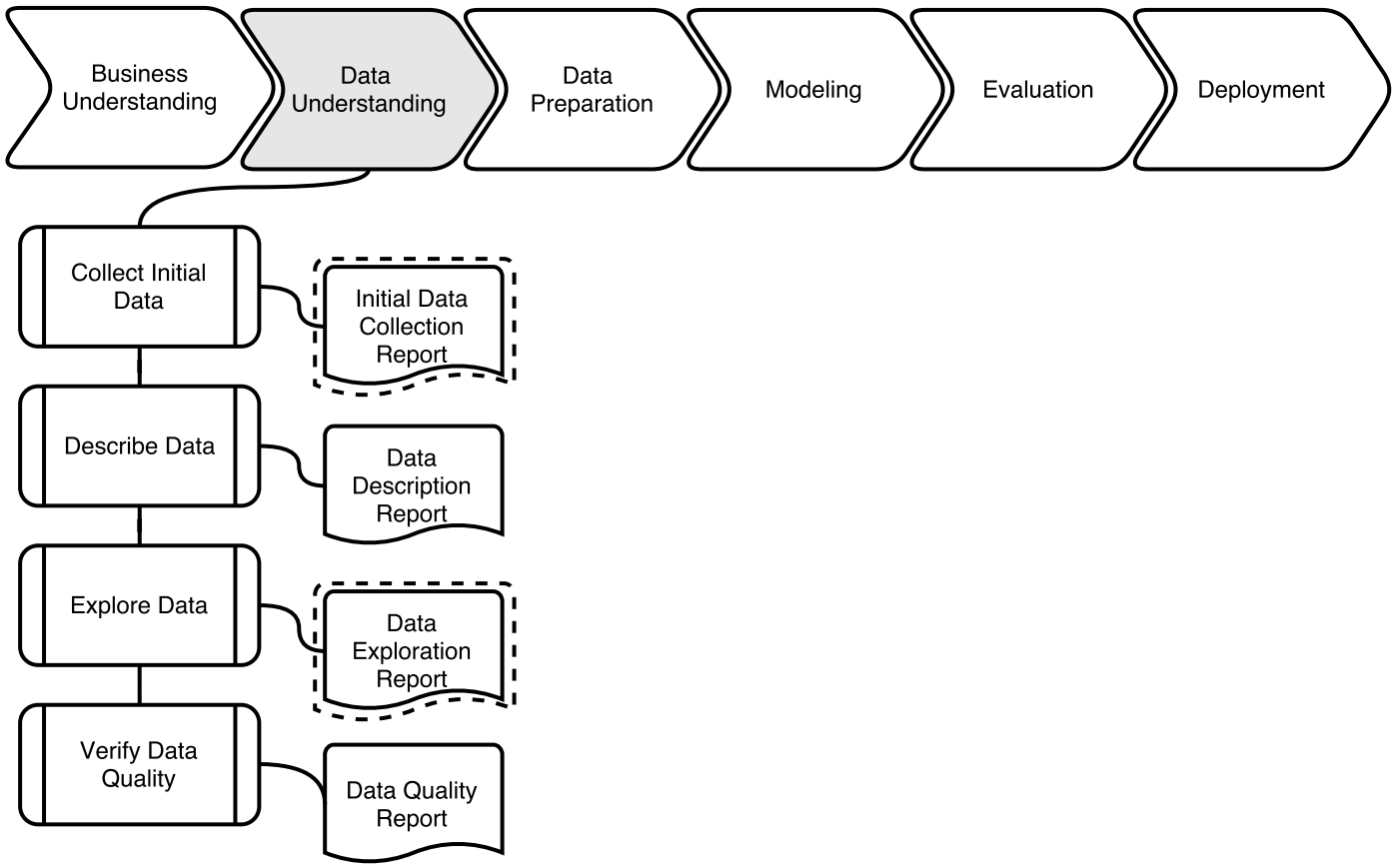

Figure 12: Phase 2. Data Understanding: tasks and activities for context-awareness

The CRISP-DM phase 2 "Data understanding" involves an initial data collection and proceeds with activities that enable you to become familiar with the data, identify data quality problems, discover first insights into the data, and/or detect interesting subsets to form hypotheses regarding hidden information.

To adapt this second phase to address the new needs, we have to enhance the initial data collection task in order to be able represent different relevant contexts. Through a further data exploration we should be also able to contribute to or refine the data description, quality reports and information about context representation, and feed into the transformation and other data preparation steps needed for further analysis.

4.2.1 Collect initial data

- **Task**: Acquire the data (or access to the data) listed in the project resources. This initial collection includes data integration if acquired from multiple data sources. Describe attributes (promising, irrelevant, ...), quantity and quality of data. Collect sufficiently rich raw data to represent possibly different relevant contexts. Collect sufficiently rich raw data to represent possibly different relevant contexts.
- **Outputs**:
  - **Initial Data Collection Report**: Describe data collected: describe attributes (promising, irrelevant, ...), quantity and quality of data and identify relevant contexts.

MoReBikeS example 5.

Initial Data Collection
A procedure to store the number of bikes in all stations every hour has been set up. The gathered data provides information about 275 bike rental stations in Valencia over a period of 2.5 years (from 01/06/2012 to 31/01/2015). For each hour in this period the data specified the median number of available bikes during that hour in each of the stations. The dataset was complemented with weather information about the same hour (temperature, relative humidity, air pressure, amount of precipitation, wind directions, maximum and mean wind speed).
The bike rental data for Valencia have been obtained from `http://biciv.com`, weather information from the Valencian Regional Ministry for the Environment (`http://www.citma.gva.es/`) and holiday information from `http://jollyday.sourceforge.net/`.

4.2.2 Describe data

- **Task**: Describe the properties of the acquired data and report on the results. This includes the amount of data (consider sampling), value types, records, fields, coding schemes, etc.
- **Outputs**:
  - **Initial Data Collection Report**: Write description report in order to share the findings about the data.

MoReBikeS example 6.

Describing Data
There are 24 given features in total which can be divided to 4 categories:

- **Facts of stations**. The facts of stations provided in the data set include the station ID, the latitude, the longitude and the number of docks in that station. All these properties for one station do not change over time.
- **Temporal information**. The timestamp of a data entry consists of eight fields: "Timestamp" in terms of seconds from the UNIX epoch, "Year", "Month", "Day", "Hour", "Weekday", "Weekhour", and "IsHoliday" which indicates whether the day is a public holiday. These features are giving overlapping temporal information, we only need a subset of them to represent a time point. The "Timestamp" is actually including information of

"Year", "Month", "Day", "Hour","Weekday" and "Weekhour", whereas "Weekday" and "Hour" also can be deduced by "Weekhour". Only "IsHoliday" is independent to any of others.

- **Weather**. This set of features include "windMaxSpeed", "windMeanSpeed", "windDirection", "temperature", "relHumidity", "airPressure", "Precipitation". One major observation of weathers is that all the values of all the seven fields share among all stations.
- **Counts and their statistics**. This set of features relates to the target value directly. First of all, "bikes 3h ago" gives the target value of the 3-hour-earlier time point at a station. The full profile features use all previous data points of the same "Weekhour" to obtain long term statistics for each "Weekhour" in each station, accordingly the short profile features only use at most four previous data points to obtain short-term statistics. The long-term statistics of the 200 old stations only have very small changes over time in contrast to the short-term ones

The target variable is "bikes" and it is a non-negative integer representing the median number of available bikes during the respective hour in the respective rental station.

4.2.3 Explore data

- **Task**: This task addresses data mining and context-aware goals through querying, visualization, and reporting techniques over the data and how they may contribute/refine the initial (business or DM) goals, data transformation/preparation.... Among others, this analysis include distribution of key attributes, looking for errors in the data, relationships between pairs or small numbers of attributes, results of simple aggregations, properties of significant subpopulations, and simple statistical analyses.
- **Outputs**:
  - **Data exploration report**: Describe results of this task including (possibly using graphs and plots) first findings, initial hypothesis, explorations about contexts, particular subsets of relevant data and attributes and their impact on the remainder of the project.

MoReBikeS example 7.

Exploring Data
A lot of work should be done in this stage in the bike scenario. Taking pieces of domain knowledge and checking whether they hold and identify interesting patterns. We can see that different stations clearly exhibit different daily patterns. Most obviously, there are stations that tend to be full in the night and emptier during the day. Essentially these are stations that are on the outer areas of the city, and the bikes are used during the day to travel into more central parts of the city. There are also stations that exhibit the opposite pattern. These stations are left empty at night, since the operators know that the will fill up during the day as people travel into the city. There are of course stations that fall between these two extremes.

4.2.4 Verify data quality

- **Task**: Examine the quality of the data: coding or data errors, missing values, bad metadata, measurement errors and other types of inconsistencies that make analysis difficult.
- **Outputs**:
  - **Initial Data Collection Report**: List and describe the results of the data quality verification (is correct?, contain errors?, missing values?, how common are they?) and list possible solutions.

MoReBikeS example 8.

Verifying Data Quality
Some of the issues encountered include missing values in the profile information about the station that could be ignored. Timepoint features also have missing values and only the timepoints with existing values are used.

4.3 Data Preparation

The CRISP-DM phase 3 "Data preparation" covers all activities needed to construct the final dataset (data that will be fed into the modelling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. In fact, it is estimated that data preparation usually takes 50-70% of a projectâĂŹs time and effort. Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modelling tools. The original CRISP-DM task "select data" had limitations for practical application in complex domains (e.g. multidimensional) since it is mostly assumed for single-table static data format. Furthermore, it lacks activities to handle data wrangling, data conversion, data sampling, data unification, etc.

Select data has been enhanced with feature extraction, resolution change and dimensionality reduction techniques to define possible attribute sets for modelling activities. Furthermore a selection of contexts and context changes relevant to the data mining goals should be done by selecting data which cover the selected contexts and changes. Enhanced constructive data preparation operations has been added to derive context-specific and context-independent attributes. The integration of data from multiple tables or records to create new records or values should be also updated with data from different contexts. Finally, data formatting for specific data mining techniques need to include the context representation.

4.3.1 Select data

- **Task**: Based upon the initial data collection conducted in the previous CRISP-DM phase, you are ready to decide on the data to be used for analysis. Note that data selection covers selection of records (rows) as well as attributes (columns) in a table.
- **Outputs**:
  - **Rationale for inclusion/exclusion**: List the data and context to be included/excluded and the reasons for these decisions.

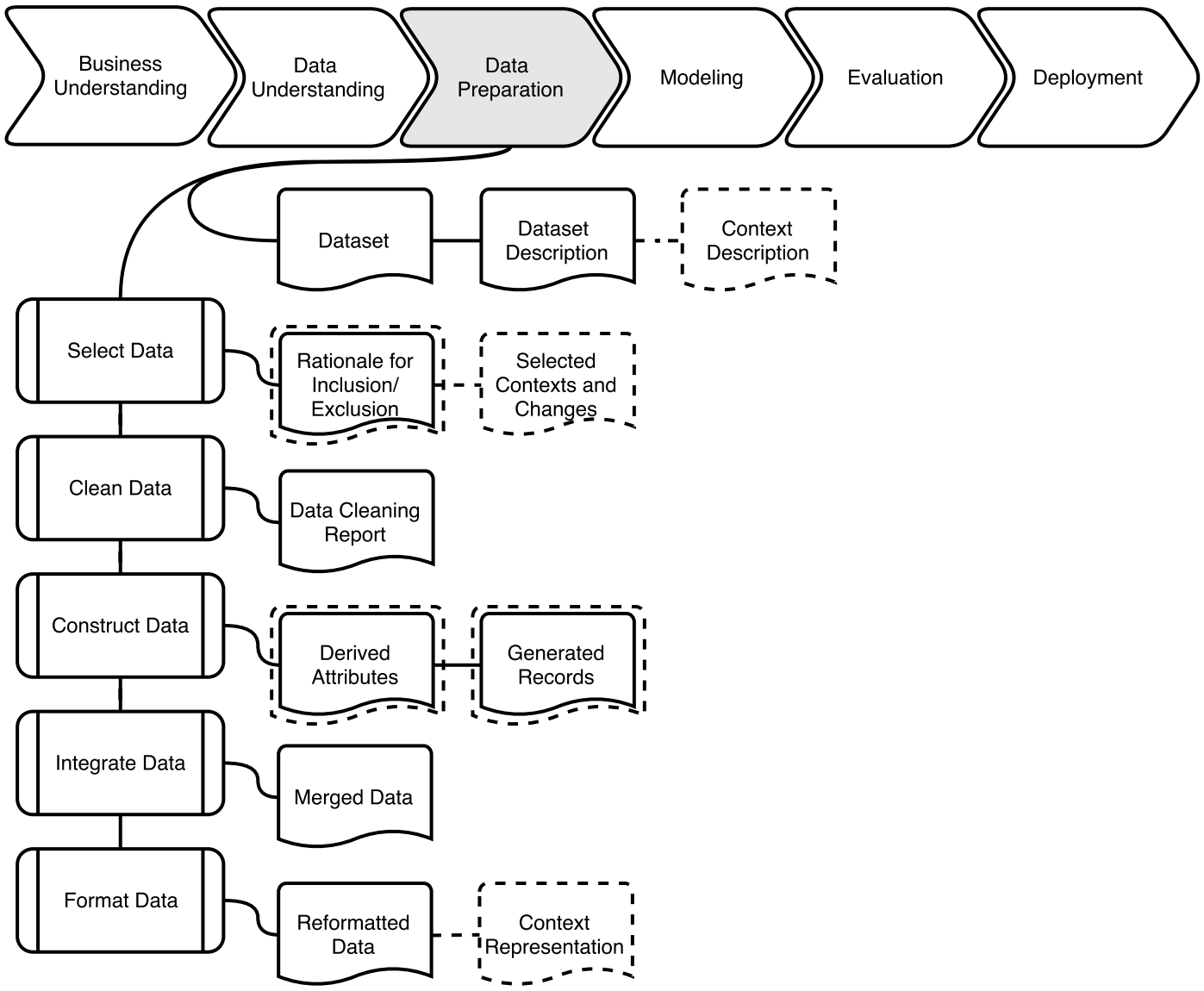

Figure 13: Phase 3. Data Preparation: tasks and activities for context-awareness

– **Selected contexts and changes**: Select contexts and context changes relevant to the data mining goals, ignore the others. Select data to cover the selected contexts and changes.

MoReBikeS example 9.

Selecting Data
Many of the decisions about which data and attributes to select have already been made in earlier phases of the data mining process. Contexts are modelled as parameters (station and timestamp) and both need to be modelled later (using all available data).

4.3.2 Clean Data

- **Task**: Clean and solve problems in the data chosen to include for the analysis. This tasks aims at raising the data quality to the level required by the selected analysis techniques.

- **Outputs**:
  - **Data Cleaning Report**: Report data-cleaning efforts (missing data, data errors, coding inconsistencies, missing data and bad metadata) for tracking alterations to the data and in order for future data mining projects to be benefited.

MoReBikeS example 10.

Selecting Data
The bike rental company uses the data cleaning process to address the problems noted in the data quality report.

- **Missing data**. The missing values are ignored in all profile calculations, i.e. only the timepoints with existing values are averaged.

4.3.3 Construct data

- **Task**: This task includes constructive data preparation operations such as the production of derived attributes or entire new records, or transformed values for existing attributes.
- **Outputs**:
  - **Derived attributes**: Derived attributes are new attributes that are constructed from one or more existing attributes in the same record. Derive context-specific and context-independent attributes.
  - **Derived attributes**: Describe the creation of completely new records. Generate new data to force context-invariance (e.g., rotated images in deep learning).

MoReBikeS example 11.

Constructing Data
Several new parameters are created to be added to the profiles of each station:

- There is one feature about the number of bikes in the station 3 hours ago: "bikes 3h ago". The profile variables are calculated from earlier available timepoints on the same station.
- The "full profile bikes" feature is the arithmetic average of the target variable "bikes" during all past timepoints with the same weekhour, in the same station.
- The "full profile 3h diffbikes" feature is the arithmetic average of the calculated feature "bikes-bikes 3h ago" during all past timepoints with the same weekhour, in the same station.
- The "short *" profile is the same as the full profiles except that it only uses past 4 timepoints with the same weekhour. If there are less than 4 such timepoints then all are used.

4.3.4 Integrate data

- **Task**: These are methods whereby information is combined from multiple sources. There are two basic methods of integrating data: merging two data sets with similar records but different attributes or appending two or more data sets with similar attributes but different records.
- **Outputs**:
  - **Merged data**: This includes: merging tables together into a new table; aggregation of data (summarising information) from multiple records and/or tables and integrating data from relevant contexts

MoReBikeS example 12.

Selecting Data
With multiple data sources (bike station' historical data, bike stations' current status, weather conditions and profile data) it is necessary to integrate all data.

4.3.5 Format data

- **Task**: This task involves checking whether certain techniques require a particular format or order to the data. Therefore syntactic modifications have to be made to the data (without changing its meaning).
- **Outputs**:
  - **Reformatted data**: Syntactic changes made to satisfy the requirements of the specific modeling tool. Examples: change the order of the attributes and/or records, add identifier, remove commas from within text fields, trimming values, etc.
  - **Context representation**: Select context representation. (How are the contexts going to be represented in the data (parametrisation; as-feature vs as-dataset)?)

MoReBikeS example 13.

Context representation
As commented in previous phases, context in this challenge is represented by means of parameters, concretely station identifier and timestamp.

4.4 Modelling

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type and this phase is usually conducted in multiple iterations. Some techniques have specific requirements on the form of data, so going back to the data preparation phase is often necessary.

A new optional branch of reframe-based subtasks and deliverables has been added for selecting the modelling technique. Therefore, we clearly differentiate between classical modelling techniques and reframing techniques. Furthermore, enhanced procedures for testing the versatile modelâĂŹs quality and validity (context plots and performance metrics) has been added. Specific reframing

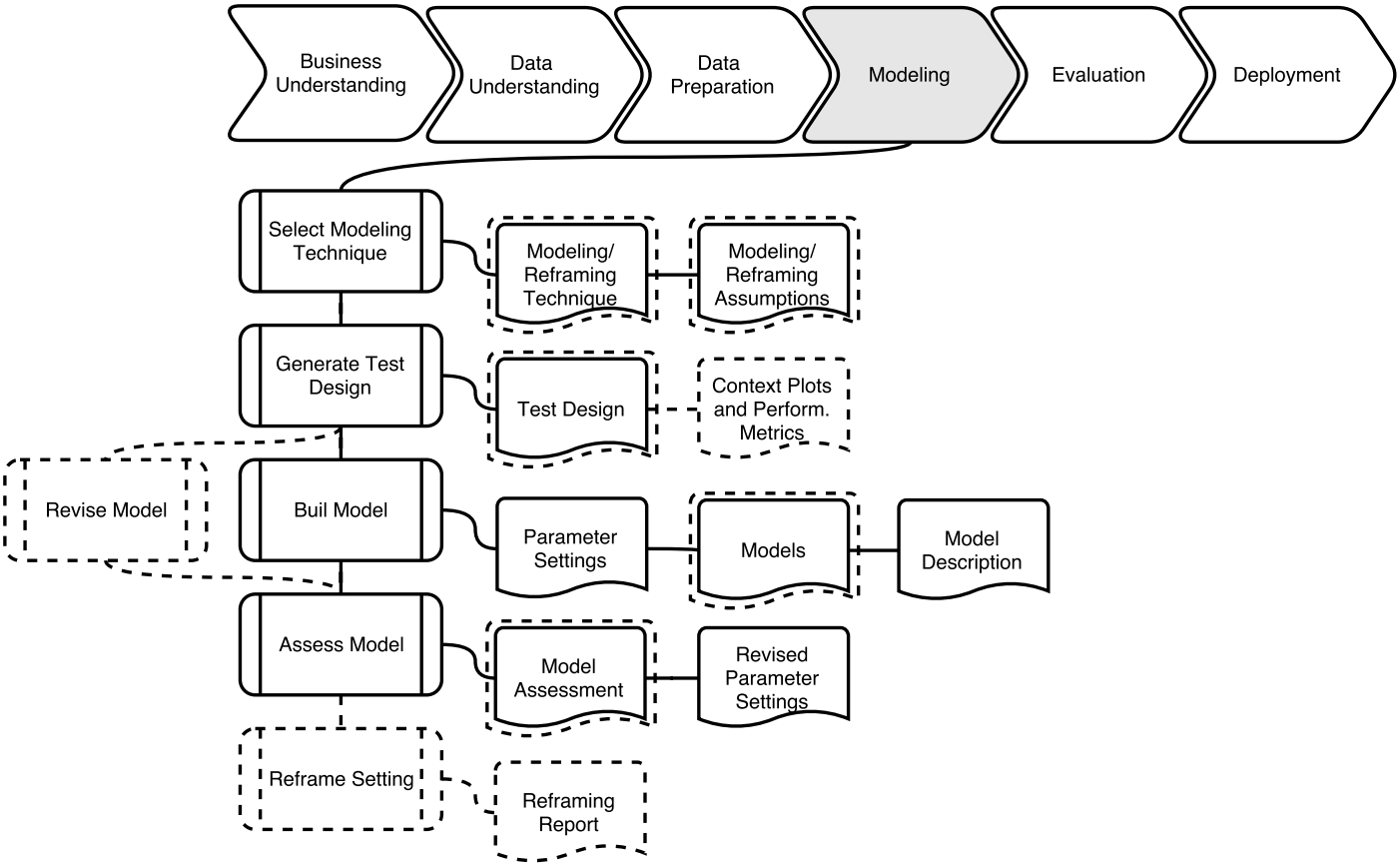

Figure 14: Phase 4. Modelling: tasks and activities for context-awareness

tools are needed to build the versatile model. A new general task âĂIJREVISE MODELâĂİ for handling model revision in incremental or lifelong learning data mining tasks. Furthermore, a new general task âĂIJREFRAME SETTINGâĂİ has been added in this phase in order to decide which type of reframing should be used (over the versatile model) depending on what aspects of the model are reusable in other contexts. This task will be performed to adapt a versatile model w.r.t. a context whenever the context changes. Finally, context-aware performance metrics are also needed to assess the versatile model.

4.4.1 Select modelling technique

- **Task**: As the first step in modelling, select the actual modeling technique that is to be used. Although it has been selected a tool during the “Business Understanding” phase, this task refers to the specific modeling technique, e.g., decision-tree building with 5.0, or neural network generation with back propagation. Determining the most appropriate model will typically be based on the data types, data mining goals (scores, patterns, clusters, versatile model. etc.)

- **Outputs**:

  - **Modeling technique**: Document the actual modeling technique that is to be used. In case context matters, Select the model and reframing couple, e.g.. scorer and score-driven or linear regression and continuous output reframing.

- **Modeling assumptions**: Many modeling techniques make specific assumptions on the data, e.g., all attributes have uniform distributions, no missing values allowed, class attribute must be symbolic etc. Record any such assumptions made.

MoReBikeS example 14.

Modeling Techniques
Because of that lack of training data, it was supposed that historical models about two-year data on the old bike stations would yield better predictions that the scarce training data. Each station is thus characterised by 6 linear models to predict the number of bikes, corresponding to 6 different subsets of the features "ikes 3h ago", "full profile bikes", "full profile 3h diffikes", "short profile bikes", "short profile 3h diff bikes" and temperature. All the subsets included "bikes 3h ago" but differed based on which profile features they used (3 options: full profiles, short profiles, or all profiles), and whether they used temperature (2 options: yes or no). We use regression modelling techniques techniques in order to handle data with outliers. We also impute Missing Values by median/mode.
Therefore , the hypothesis made was that the closest old stations to the target stations were most capable to predict future use of those new stations given the different models for the other 200 stations. For that reason, distance seems to be a crucial point in weighting the predictions of the given models.

4.4.2 Generate test design

- **Task**: Before we actually build a model, we should consider how the modelâĂŹs results will be tested. Therefore we need to generate a procedure or mechanism to test the modelâĂŹs quality and validity (describing the criteria for goodness of a model (i.e., error rate) and defining the data on which these criteria will be tested.
- **Outputs**:
  - **Test design**: Describe the intended plan (i.e., how to divide the available dataset) for training, testing and evaluating the models.
  - **Context plot and performance metrics**: Decide how the context changes can be evaluated (e.g, by using artificial data). Identify proper metrics to evaluate reframing efficiency.

MoReBikeS example 15.

Test Design
As already commented, the stations are first splitted randomly into 200 training stations and 75 test stations. The time period was splitted into training period (01/06/2012 to 31/10/2014) and three-months test period (01/11/2014 to 31/01/2015). The last month of the training period (01/10/2014 to 31/10/2014) we referred to as the deployment period. We trained 6 different linear models (more details in the previous task) for each of the 200 training stations on the training period. The participants are provided with the trained models, with the data from the one-month deployment period for all 200+75 stations, and with the data from the training period from 10 training stations out of 200.
The criteria by which the models are assessed is the Mean Absolute Error (MAE) in three-

month test period across 50 test stations, with different forecasting windows, grouped by length of history and perhaps some meta-information about the station.

4.4.3 Build Model

- **Task**: Run the modelling tool on the prepared dataset to create one or more models.
- **Outputs**:
  - **Parameter settings**: Most modeling techniques have a a large number of parameters that can be adjusted. List the parameters and their chosen value, along with the rationale for the choice of parameter settings.
  - **Models**: These are the actual models produced by the modeling tool, not a report.
  - **Model description**: Describe the resultant model. Report on the results of a model and any meaningful conclusion, document any difficulties or inconsistencies encountered with their meanings.

MoReBikeS example 16.

Model Building
In choosing the models to be reused there is a good range of approaches: perform an analysis to select for each test station one model out of the given 1200, and used that model for prediction. Select multiple models and averaged over these. Use a weighted average of model. Finally, and following the retraining approach, it should be decided not to use the given model and trained new models.

- **Reframe** version: a possible solution to the problem consists of combining the predictions of the K nearest stations among the old stations (1:200) to the target stations (201:275) using the weighted arithmetic mean. On one hand, these predictions are calculated applying the best model—in terms of MAE—for each old station (1:200). On the other hand, the K nearest neighbours were obtained by comparing each target stations (201:275) to all the old stations (1:200) in terms of the Euclidean distance between them. Then, the K closest old stations to one target station were selected as its K nearest neighbours. In doing so for every target station (201:275), their K nearest neighbours were discovered among the old stations (1:200). The Euclidean distance between the target station and its neighbours is used to weight the influence of their predictions on the final prediction. Finally, this summation was divided by the sum of the k Euclidean distances from each neighbour (among the K nearest neighbours) to the target station on the test data. In doing so, the final prediction value was obtained from k predictions taken into account in a different importance according to their proximity to the target station.
- **Retraining** version: it consists on using the data of the roughly 2.5 year long period between 2012 and 2014 for 10 docking stations in the city of Valencia as well as the one month partial training data provided for 190 other stations throughout the city.

4.4.4 Revise Model

- **Task**: Once we have built a model and as a result of an incremental learning or lifelong learning, the model needs to be revised(patched or extended) because of some novelty or inconsistency of the new data is detected with respect to the existing model. This can be extended to context changes, provided we can determine when the context has changed significantly to deserve a revision process.

4.4.5 Assess Model

- **Task**: For each model under consideration, we have to interpret them and make a methodical assessment according to the data mining success criteria, and the desired test design. Judge and discuss the the success of the application of modelling and discovery techniques technically. Rank the models used.
- **Outputs**:
  - **Model assessment**: Summarize results of this task by using evaluation chars, analysis nodes, cross-validation charts, etc.; list qualities of generated models (e.g., in terms of accuracy) and rank their quality in relation to each other. In context-aware tasks, compare with different scenarios, in particular retraining.
  - **Revised parameter settings**: According to the model assessment, revise parameter settings and tune them for the next run in the Build Model task. Iterate model building and assessment until you strongly believe that you found the best model(s). Document all such revisions and assessments.

MoReBikeS example 17.

MoReBikeS Example—Model Assessment
The reframing solution, which selects multiple models and averages over these, is better than retraining.

4.4.6 Reframe setting

- **Task**: Which type of reframing technique should be used depending on what aspects of the model are reusable in other contexts? Taking into account the particular deployment context (if known), we distinguish three different kinds of reframing (which can be combined): *output*, *input* and *structural* reframing. Thus, where a conventional, non-versatile model captures only such information as is necessary to deal with test instances from the same context, a versatile model captures additional information that, in combination with reframing, allows it to deal with test instances from a larger range of contexts.
- **Outputs**:
  - **Kind of reframing**: Describe the kind of reframing (output, input or structural) to be applied over the versatile model.

MoReBikeS example 18.

Reframe setting
In choosing the models to be reused it has to be decided the criteria for model suitability for a given test station. These included: performance of the model in the test station during the deployment period; distance between the test station and the station of the model's origin; similarity between the time-series of the stations during the deployment period; and several combinations of these.

4.5 Evaluation

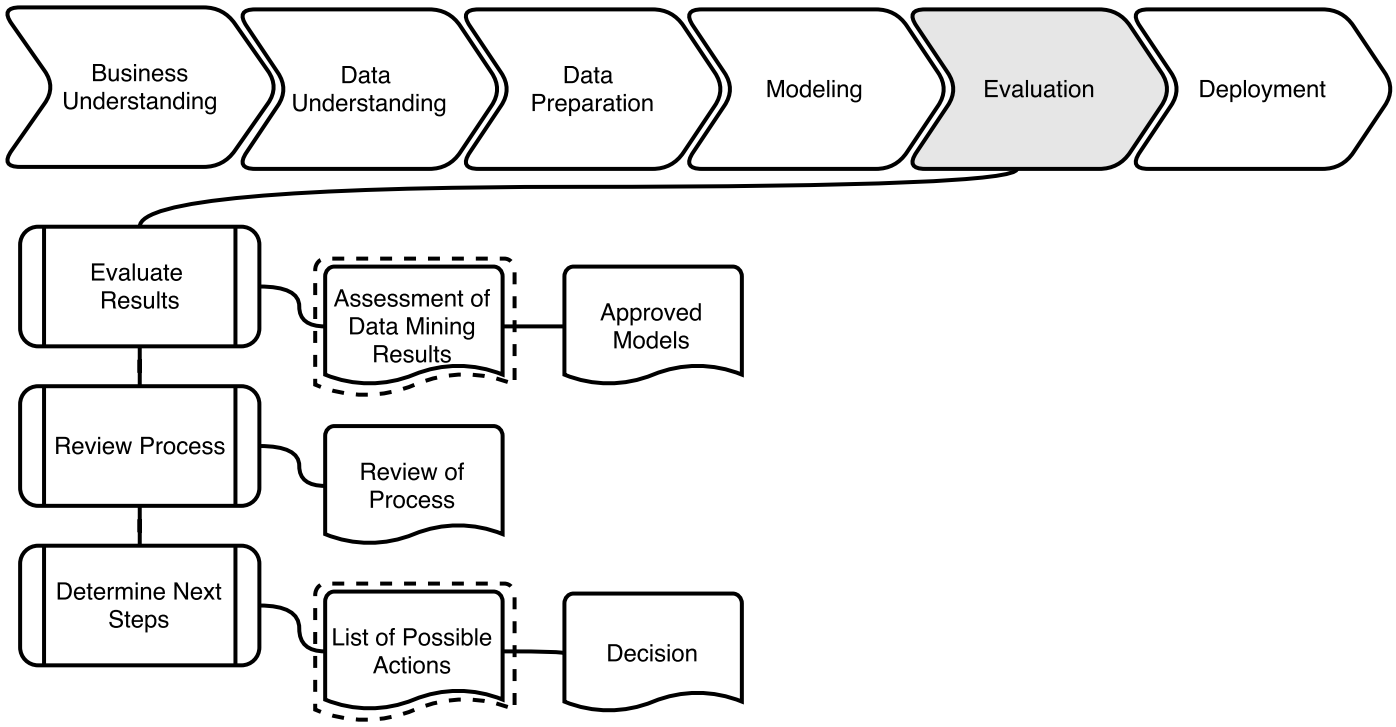

Figure 15: Phase 5. Evaluation: tasks and activities for context-awareness

Once you have built a model (or models) that, according to the evaluation task in the previous phase, appears to have high quality from a data analysis perspective, it is important to thoroughly evaluate it (perhaps going through the previous phases) to be certain the model properly achieves the business objectives. Therefore, this step requires a clear understanding of the stated business goals. A key objective is to determine how well results answer your organizationâĂŹs business goals and whether there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

Regarding context-awareness, in this phase we need an enhanced task for assessing whether the versatile model meets the business objectives in all the relevant contexts where they are to be deployed. Furthermore, we need to decide whether the versatile model is able to be reused and adapted to the deployment data or not.

4.5.1 Evaluate results

- **Task**: Unlike the previous evaluation steps which dealt with factors such as the accuracy and generality of the model, in this step we need to assess the degree to which the model meets the business objectives and, thus, this step requires a clear understanding of the stated business goals. We need to determine if there is some business reason why this model is deficient, if results are stated clearly, if there are novel or unique findings that should be highlighted, if results raised additional questions, etc.
- **Outputs**:
  - **Assessment of data mining results with respect to business success criteria**: Summarize assessment results in terms of business success criteria, interpret the data mining results, check the impact of result for initial application goal in the project, see if the discovered information is novel and useful, rank the results, state conclusions, check whether results cover all contexts relevant for the business success criteria, etc.
  - **Approved models**: Select those (versatile) models which, after the previous assessment with respect to business success criteria, meet the the selected criteria.

MoReBikeS example 19.

Evaluating Results
Given a new rental station (deployment context), it is conceivable that there might be some rental stations that are more similar to this station in terms of the daily usage patterns. Following this idea, the proposed here methods find the closest stations in terms of distance. The overall results indicate that the deployed methods are quite simple and easy to apply and can achieve a good performance when used to know which stations are more likely to be empty or full soon.

- **New Questions**. The most important questions to come out of the study are: How often the stations remain empty or full because of bad predictions? How much time is wasted in carrying bikes around because of bad predictions? Can we use different evaluation measures in modelling to achieve better results? How often do we need to retrain models?

4.5.2 Review Process

- **Task**: Extra time for reflection on the successes and weaknesses of the process just completed. Although the resulting models appear to be satisfactory and to satisfy business needs, it would be appropriate to do a more thorough review of the whole data mining process seeking for overlooked tasks and quality assurance issues. We should summarise activities and decisions made in each phase learning thus from your experience so that future data mining projects will be more effective.
- **Outputs**:
  - **Review of process**: Summarize the process review and all the activities and decisions for each phase. Give hints for activities that have been missed and/or should be repeated.

MoReBikeS example 20.

MoReBikeS Example—Review Report
As a result of reviewing the process of the initial data mining project, the bike rental company has developed a greater appreciation of the interrelations between steps an its inherent "backtracking" nature. Furthermore, the company has learn that model reuse between similar stations is appropriate when historical data is not provided or does not exists.

4.5.3 Determine next steps

- **Task**: Depending on the results of the reviewing the process of the initial data mining project, the project team decides how to proceed. The team decides whether (a) to continue to the deployment phase, (b) go back and refine or replace your models thus initiating further iterations, or (c) set up new data mining projects. This task includes analyses of remaining resources and budget, which may influence the decisions. If the results satisfactorily meet your data mining and business goals, start the deployment phase.
- **Outputs**:
  - **List of possible actions**: List possible further actions along with the reasons for and against each option: analyse potential for deployment and improvement(for each result obtained), recommend alternative following phases, refine the whole process, etc.
  - **Decision**: Describe the decision made: rank alternatives, document reasons for the choice and how to proceed along with the rationale.

MoReBikeS example 21.

Next Steps
The bike rental company is fairly confident of both the accuracy and relevancy of the project results and so is continuing to the deployment phase.

4.6 Deployment

Creation of the model is generally not the end of the project and deployment is the process of using the discovered insights to make improvements (or changes) within your organization. Even if the results may not be formally integrated into your information systems, the knowledge gained will undoubtedly be useful for planning and making marketing decisions.This phase often involves planning and monitoring the deployment of results or completing wrap-up tasks such as producing a final report and conducting a project review.

Regarding context-awareness data mining tasks, in this phase we need to determine in what way the versatile model (or the pull of models) is to be kept, used, evaluated and maintained for a long-term use. Furthermore, we may need to monitor the possible change of the context distribution or check whether its range is the same as expected. If not, we may need to revaluate some models for a new distribution of contexts thus going back to previous phases/tasks.

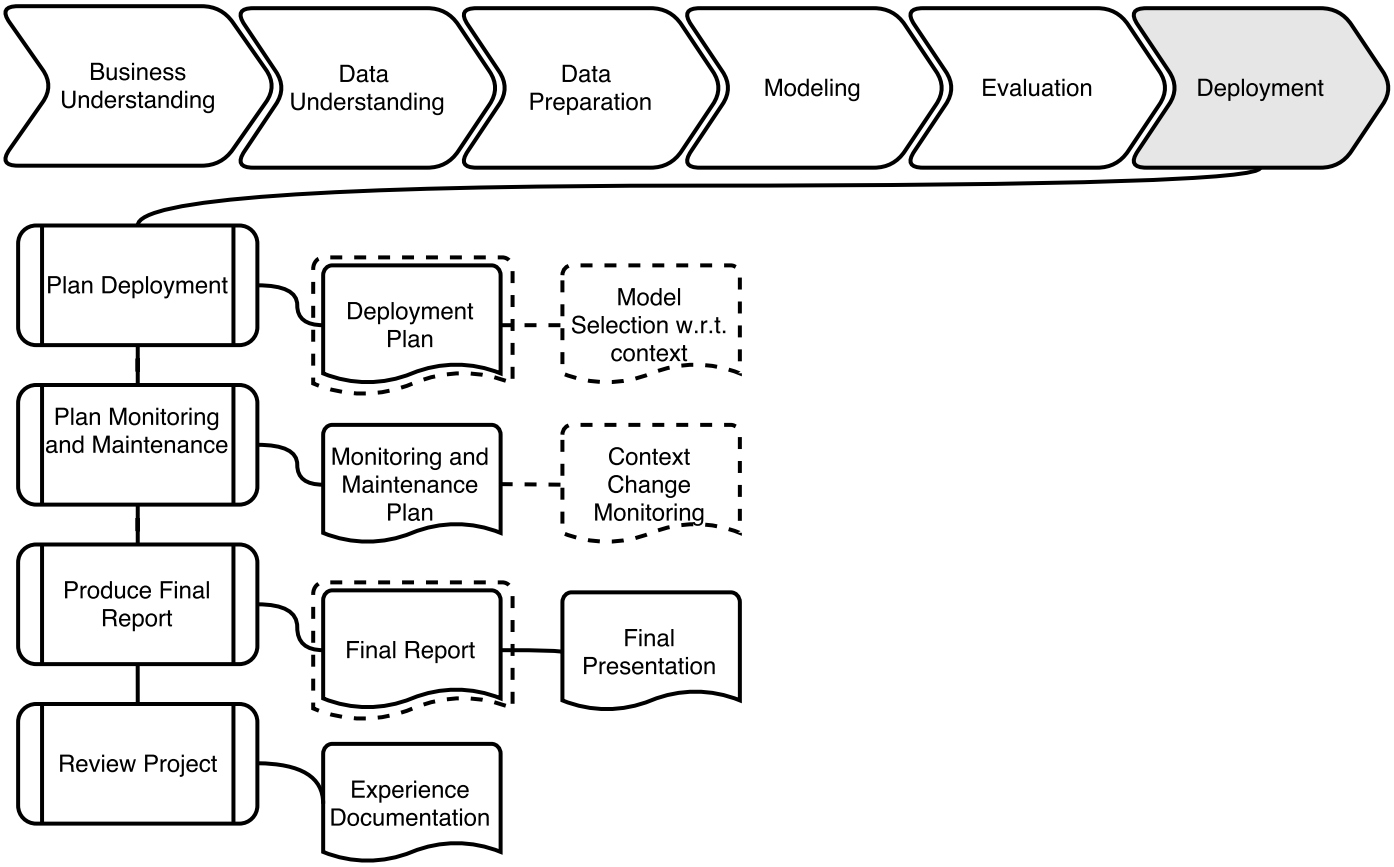

Figure 16: Phase 6. Deployment: tasks and activities for reframing.

4.6.1 Plan deployment

- **Task**: Depending on the results of the reviewing the process of the initial data mining project, the project team decides how to proceed. The team decides whether (a) to continue to the deployment phase, (b) go back and refine or replace your models thus initiating further iterations, or (c) set up new data mining projects. This task includes analyses of remaining resources and budget, which may influence the decisions. If the results satisfactorily meet your data mining and business goals, start the deployment phase.
- **Outputs**:
  - **Deployment plan**: This task takes the evaluation results and determines a strategy for deployment. If a general procedure has been identified to create the relevant model(s) and integrate within your database systems. This procedure is documented (step-by-step plan and integration) here for later deployment (including technical details, benefits of monitoring, deployment problems, etc.). Furthermore, create a plan to disseminate the relevant information to strategy makers
  - **Model selection w.r.t. the context**: Determine the pace in which context values are captured or estimated. Determine how the pool of models is going to be kept and selected according to context.

MoReBikeS example 22.

Deployment Planning
Can we use different evaluation measures in modelling to achieve better results? How often do we need to retrain models?.

4.6.2 Plan deployment

- **Task**: Since your data mining work may be ongoing, monitoring and maintenance are important issues. In those cases the model(s) will likely need to be evaluated periodically to ensure its effectiveness and to make continuous improvements.
- **Outputs**:
  - **Monitoring and maintenance plan**: Summarize monitoring and maintenance strategy: factors or influences need to be tracked, validity and accuracy of each model, expiration issues,etc.
  - **Context change Monitoring**: Determine the pace in which context values are captured or estimated. Determine how the pool of models is going to be kept and selected according to context.

MoReBikeS example 23.

Deployment Planning
The company had to decide on the criteria for model suitability for a given test station. In choosing the models to be reused in each of the new bike stations to be allocated, the bike rental company will select the set provided models (from existing stations) that have already been proved to be similar to the new station.

4.6.3 Produce final report

- **Task**: At the end of the project, the project team writes up a final report to communicate the results
- **Outputs**:
  - **Final report**: Final report where all the threads are brought together. It should include a thorough description of the original business problem, the process used to conduct data mining, how well initial data mining goals have been met, which (versatile) models are reused again and again, budget and costs (cost of reframing? And retraining? How significant has context been?)), deviations from the original plan, summary of data mining results, overview of the deployment process, recommendations and insights discovered, etc.
  - **Final presentation**: Determine the pace in which context values are captured or estimated. Determine how the pool of models is going to be kept and selected according to context.

4.6.4 Produce final report

- **Task**: This is the final step of the CASP-DM methodology. In it we assess what went right and what went wrong (and need to be improved), the final impressions, lessons learned, etc.
- **Outputs**:
  - **Experience documentation**: Summarize important experiences made during the project (overall impressions, pitfalls, misleading approaches, etc.). Have contexts been well identified? How much model reuse has been performed? Were the models sufficiently flexible to be reframed? Should we change the definition of context? Can we make more versatile models?

5 Discussion

Data mining is a discipline with strong technical roots in statistics, machine learning and information systems. The advance in techniques, tools and platforms, jointly with the increase of the availability of data and the higher complexity of projects and teams, has been so significant in the past decade that methodological issues are becoming more important to harness all this potential in an efficient way. The perspective of data science, where data mining goals are more data-oriented than business-oriented in a more classical direct data mining process may suggest that rigid methodologies cannot cope with the variability of problems, which have to be adjusted to related scenarios very frequently, in terms of changes of data, goals, resolution, noise or utility functions.

In contrast, we have advocated here that successful methodologies, such as CRISP-DM, can play this role if they become less rigid and accommodate the idea the variability of the application in a more systematic way. The notion of context, its identification and parametrisation, is a general way to anticipate all these changes and consider them from the very beginning. This is why CASP-DM tries to extend CRISP-DM to make this possible. The explicit existence of activity and tasks specifically designed for this context identification and handling ensures that companies and practitioners will not overlook this important aspect and will plan data mining projects in a more robust way, where data transformation and model construction can be reused and not jettisoned whenever any contextual thing changes. We have illustrated how CASP-DM goes through these context issues with some real examples.

CASP-DM not only considers context-awareness in the whole process, but is backward compatible with CRISP-DM, the most common methodology in data mining. This means that CRISP-DM users can adopt CASP-DM immediately and even complement their existing projects with the context-aware bits, making them more versatile. In order to do this transition from CRISP-DM to CASP-DM, it is also important to have a stable platform and community where CASP-DM documents, phases and planning tools can be integrated and located for data mining practitioners. For instance, it is hard to find the CRISP-DM documentation, as nobody is maintaining it any more. To take that reference role, we have set up a community around `www.casp-dm.org`, where data mining practitioners can find information about CRISP-DM and CASP-DM, but also about context-awareness and other related areas such as reframing and domain adaptation. It is also our intention to associate a working group with this initiative, so that CASP-DM can also evolve with the new methodological challenges of data mining.

References

Abowd, G. D., Dey, A. K., Brown, P. J., Davies, N., Smith, M., and Steggles, P. (1999). Towards a better understanding of context and context-awareness. In *Handheld and ubiquitous computing*, pages 304–307. Springer.

Anand, S. S. and Büchner, A. G. (1998). *Decision support using data mining*. Financial Times Management.

Anand, S. S., Patrick, A., Hughes, J. G., and Bell, D. A. (1998). A data mining methodology for cross-sales. *Knowledge-Based Systems*, 10(7):449–461.

Angluin, D. and Laird, P. (1988). Learning from noisy examples. *Machine Learning*, 2(4):343–370.

Bareinboim, E. and Pearl, J. (2012). Transportability of causal effects: Completeness results. In *AAAI*.

Bi, J. and Bennett, K. P. (2003). Regression error characteristic curves. In *Twentieth International Conference on Machine Learning (ICML-2003). Washington, DC*.

Blanco-Vega, R., Ferri, C., Hernández-Orallo, J., and Ramírez-Quintana, M. J. (2006). Estimating the class probability threshold without training data. *ROC Analysis in Machine Learning*, page 9.

Brachman, R. J. and Anand, T. (1996). Advances in knowledge discovery and data mining. chapter The Process of Knowledge Discovery in Databases, pages 37–57. American Association for Artificial Intelligence, Menlo Park, CA, USA.

Brunk, C., Kelly, J., and Kohavi, R. (1997). Mineset: An integrated system for data mining. In *KDD*, pages 135–138.

Buchner, A. G., Mulvenna, M. D., Anand, S. S., and Hughes, J. G. (1999). An internet-enabled knowledge discovery process. In *Proceedings of the 9th international database conference, Hong Kong*, volume 1999, pages 13–27.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., and Zanasi, A. (1998). *Discovering data mining: from concept to implementation*. Prentice-Hall, Inc.

Caruana, R. (1997). Multitask learning. *Machine Learning*, 28(1):41–75.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). Crisp-dm 1.0 step-by-step data mining guide.

Chow, C. (1970). On optimum recognition error and reject tradeoff. *Information Theory, IEEE Transactions on*, 16(1):41–46.

Cios, K. J. and Kurgan, L. A. (2005). Trends in data mining and knowledge discovery. In *Advanced techniques in knowledge discovery and data mining*, pages 1–26. Springer.

Cios, K. J., Teresinska, A., Konieczna, S., Potocka, J., and Sharma, S. (2000). A knowledge discovery approach to diagnosing myocardial perfusion. *Engineering in Medicine and Biology Magazine, IEEE*, 19(4):17–25.

Debuse, J., de la Iglesia, B., Howard, C., and Rayward-Smith, V. (2001). Building the kdd roadmap. In *Industrial Knowledge Management*, pages 179–196. Springer.

Drummond, C. and Holte, R. (2006). Cost Curves: An Improved Method for Visualizing Classifier Performance. *Machine Learning*, 65:95–130.

Edelstein, H. A. (1998). *Introduction to data mining and knowledge discovery*. Two Crows.

Elkan, C. (2001). The foundations of Cost-Sensitive learning. In *IJCAI-01*, pages 973–978.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996a). The kdd process for extracting useful knowledge from volumes of data. *Commun. ACM*, 39(11):27–34.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996b). Advances in knowledge discovery and data mining.

Ferri, C., Hernández-Orallo, J., and Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1):27–38.

Flach, P. (2010). ROC analysis. In *Encyclopedia of Machine Learning*, pages 869–875. Springer.

Flach, P., Blockeel, H., Ferri, C., Hernández-Orallo, J., and Struyf, J. (2003). Decision support for data mining. In *Data Mining and Decision Support*, pages 81–90. Springer.

Flach, P., Hernández-Orallo, J., and Ferri, C. (2011). A coherent interpretation of AUC as a measure of aggregated classification performance. In *ICML*.

Frénay, B. and Verleysen, M. (2013). Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5).

Gertosio, C. and Dussauchoy, A. (2004). Knowledge discovery from industrial databases. *Journal of Intelligent Manufacturing*, 15(1):29–37.

Giraud-Carrier, C., Vilalta, R., and Brazdil, P. (2004). Introduction to the special issue on meta-learning. *Machine learning*, 54(3):187–193.

Hand, D. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine learning*, 77(1):103–123.

Harry, M. J. (1998). Six sigma: a breakthrough strategy for profitability. *Quality progress*, 31(5):60.

Hernández-Orallo, J. (2013). ROC curves for regression. *Pattern Recognition*, 46(12):3395–3411.

Hernández-Orallo, J., Ferri, C., Lachiche, N., Martínez-Usó, A., and Ramírez-Quintana, M. J. (2015). Binarised regression tasks: methods and evaluation metrics. *Data Mining and Knowledge Discovery*, pages 1–43.

Hernández-Orallo, J., Ferri, C., Lachiche, N., Martínez-Usó, A., and Ramírez-Quintana, M. J. (2016). Binarised regression tasks: methods and evaluation metrics. *Data Mining and Knowledge Discovery*, 30(4):848–890.

Hernández-Orallo, J., Flach, P., and Ferri, C. (2011). Brier curves: a new cost-based visualisation of classifier performance. In *ICML*.

Hernández-Orallo, J., Flach, P., and Ferri, C. (2012a). A unified view of performance metrics: Translating threshold choice into expected classification loss. *JMLR*, 13:2813–2869.

Hernández-Orallo, J., Flach, P., and Ferri, C. (2012b). A unified view of performance metrics: Translating threshold choice into expected classification loss. *Journal of Machine Learning Research*, 13:2813–2869.

Hernández-Orallo, J., Flach, P., and Ferri, C. (2013). ROC curves in cost space. *Machine Learning*, 93(1):71–91.

Hernández-Orallo, J., Usó, A. M., Prudêncio, R. B. C., Kull, M., Flach, P. A., Ahmed, C. F., and Lachiche, N. (2016). Reframing in context: A systematic approach for model reuse in machine learning. *AI Commun.*, 29(5):551–566.

Jiang, J. (2008). A literature survey on domain adaptation of statistical classifiers. *URL: http://sifaka. cs. uiuc. edu/jiang4/domainadaptation/survey*.

Khreich, W., Granger, E., Miri, A., and Sabourin, R. (2012). A survey of techniques for incremental learning of HMM parameters. *Information Sciences*, 197:105–130.

Kull, M. and Flach, P. (2014). Patterns of dataset shift. In *Ws. on Learning over Multiple Contexts at ECML2014 (LMCE)*.

Kull, M. and Hernández-Orallo, J. (2015). Missing values on purpose: Model selection and reframing with attribute and prediction costs. *submitted*.

Kull, M., Lachiche, N., and Martınez-Usó, A. (2015a). Morebikes-model reuse with bike rental station data.

Kull, M., Lachiche, N., and Usó, A. M. (2015b). Model reuse with bike rental station data (preamble). In *Proceedings of the ECML/PKDD 2015 Discovery Challenges co-located with European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2015), Porto, Portugal, September 7-11, 2015.*

Lo, H.-Y., Wang, J.-C., Wang, H.-M., and Lin, S.-D. (2011). Cost-sensitive multi-label learning for audio tag annotation and retrieval. *Multimedia, IEEE Transactions on*, 13(3):518–529.

Mariscal, G., Marban, O., and Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, 25(02):137–166.

Martínez-Usó, A. and Hernández-Orallo, J. (2015). Multidimensional prediction models when the resolution context changes. In *ECML*.

Martínez-Usó, A., Hernández-Orallo, J., Ramírez-Quintana, M. J., and Plumed, F. M. (2015). *Pentaho + R: An Integral View for Multidimensional Prediction Models*, pages 234–244. Springer International Publishing.

Metz, C. E. (1978). Basic principles of ROC analysis. In *Seminars in nuclear medicine*, volume 8,4, pages 283–298. Elsevier.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., and Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530.

Moyle, S. and Jorge, A. (2001). Ramsys-a methodology for supporting rapid remote collaborative data mining projects. In *ECML/PKDD 2001 Workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning: Internal SolEuNet Session*, pages 20–31.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359.

Pietraszek, T. (2007). On the use of ROC analysis for the optimization of abstaining classifiers. *Machine Learning*, 68(2):137–169.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2009). *Dataset shift in machine learning*. The MIT Press.

Raedt, L. D. (1992). *Interactive Theory Revision: An Inductive Logic Programming Approach*. Academic Press.

Richards, B. L. and Mooney, R. J. (1991). First-order theory revision. In *ML*, pages 447–451.

SAS (2005). Semma data mining methodology. `http://www.sas.com/technologies/analytics/datamining/miner/semma.html`.

Scheirer, W. J., de Rezende-Rocha, A., Sapkota, A., and Boult, T. E. (2013). Toward open set recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(7):1757–1772.

Thrun, S. (1996). Is learning the n-th thing any easier than learning the first? *Advances in neural information processing systems*, pages 640–646.

Thrun, S. and Pratt, L. (2012). *Learning to learn*. Springer Science & Business Media.

Torrey, L. and Shavlik, J. (2009). Transfer learning. *Handbook of Research on Machine Learning Applications*, 3:17–35.

Tortorella, F. (2005). A ROC-based reject rule for dichotomizers. *Pattern Recognition Letters*, 26(2):167–180.

Turney, P. (2000). Types of cost in inductive concept learning. *Canada National Research Council Publications Archive*.

Vanderlooy, S., Sprinkhuizen-Kuyper, I., and Smirnov, E. (2006). An analysis of reliable classifiers through ROC isometrics. In *Proceedings of the ICML 2006 Ws. on ROC Analysis (ROCML 2006), Pittsburgh, USA, June*, volume 29, pages 55–62.

Xu, Z., Kusner, M. J., Weinberger, K. Q., Chen, M., and Chapelle, O. (2014). Classifier cascades and trees for minimizing feature evaluation cost. *JMLR*, 15:2113–2144.